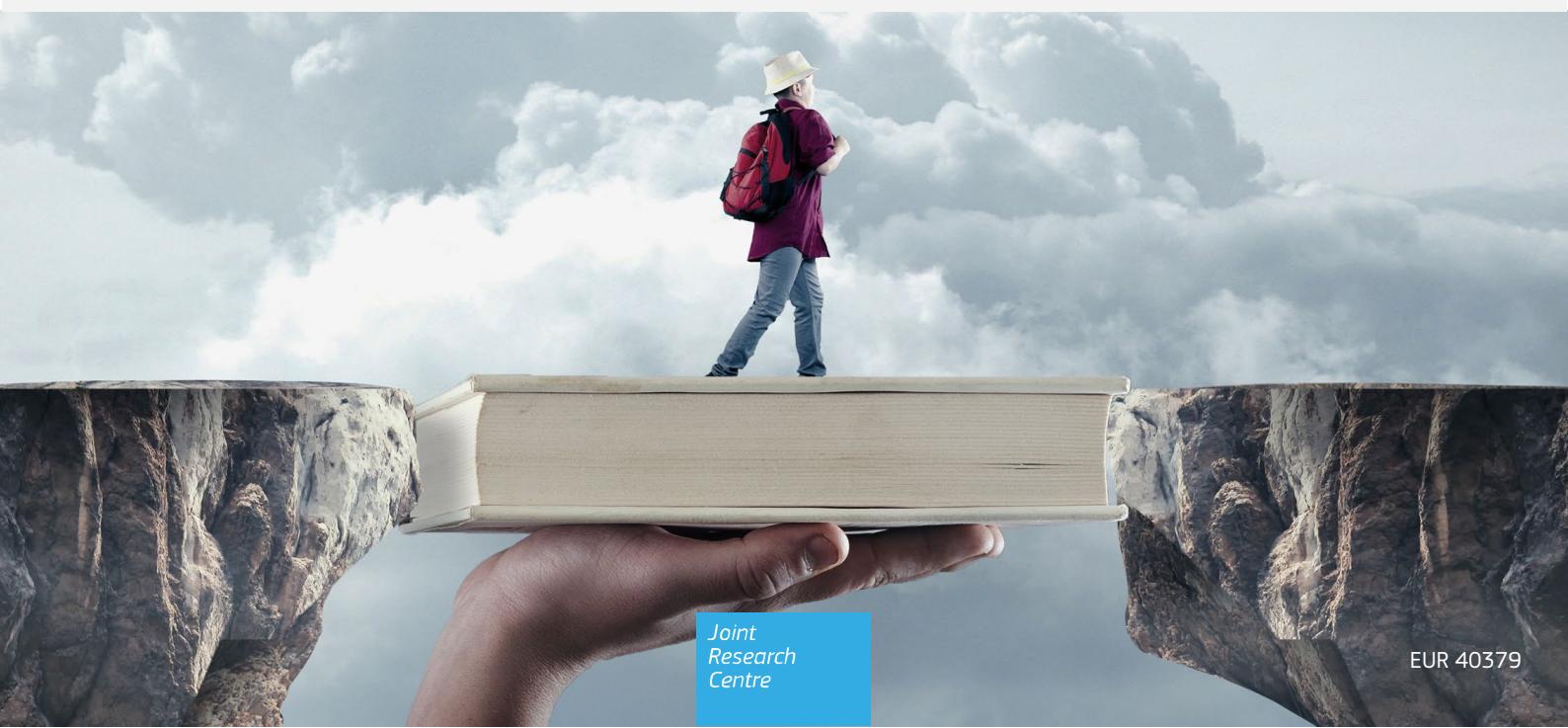




# AI evidence pathway for operationalising trustworthy AI in health: An ontology unfolding ethical principles into translational and fundamental concepts

Griesinger, C.B., Reina, V., Panidis, D., Chassaigne, H.

2025



This document is a publication by the Joint Research Centre (JRC), the European Commission's science and knowledge service. It aims to provide evidence-based scientific support to the European policymaking process. The contents of this publication do not necessarily reflect the position or opinion of the European Commission. Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use that might be made of this publication. For information on the methodology and quality underlying the data used in this publication for which the source is neither Eurostat nor other Commission services, users should contact the referenced source. The designations employed and the presentation of material on the maps do not imply the expression of any opinion whatsoever on the part of the European Union concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

### Contact information

Name: Claudius B. Griesinger

Address: Directorate F – Health and Food, Unit F.2 - Technologies for Health, Via E. Fermi 2749, I-21027 Ispra (VA), Italy

Email: [claudius.griesinger@ec.europa.eu](mailto:claudius.griesinger@ec.europa.eu) ; Tel.: +39 0332 78 6726

### EU Science Hub

<https://joint-research-centre.ec.europa.eu>

JRC140726

EUR 40379

PDF ISBN 978-92-68-29680-6 ISSN 1831-9424 doi:10.2760/8107037 KJ-01-25-369-EN-N

Luxembourg: Publications Office of the European Union, 2025

© European Union, 2025



The reuse policy of the European Commission documents is implemented by the Commission Decision 2011/833/EU of 12 December 2011 on the reuse of Commission documents (OJ L 330, 14.12.2011, p. 39). Unless otherwise noted, the reuse of this document is authorised under the Creative Commons Attribution 4.0 International (CC BY 4.0) licence (<https://creativecommons.org/licenses/by/4.0/>). This means that reuse is allowed provided appropriate credit is given and any changes are indicated.

For any use or reproduction of photos or other material that is not owned by the European Union permission must be sought directly from the copyright holders.

- Cover page illustration, ©allvision - stock.adobe.com (#240951973)

How to cite this report: Griesinger, C.B., Reina, V., Panidis, D. and Chassaigne, H., *AI evidence pathway for operationalising trustworthy AI in health: Ontology of ethical principles, translational concepts and fundamental concepts*, Publications Office of the European Union, Luxembourg, 2025, <https://data.europa.eu/doi/10.2760/8107037>, JRC140726.

## Contents

→ *For a detailed table of contents, see Annex 2.*

|   |            |
|---|------------|
| <b>Abstract .....</b>   | <b>2</b>   |
| <b>How to navigate this document.....</b>   | <b>3</b>   |
| <b>Foreword.....</b>  | <b>4</b>   |
| <b>Executive summary .....</b>  | <b>5</b>   |
| <b>Acknowledgements .....</b>   | <b>9</b>   |
| Authors.....  | 9          |
| Author contributions .....  | 9          |
| Use of artificial intelligence .....  | 9          |
| <b>1     Introduction .....</b>   | <b>10</b>  |
| 1.1     About this ontology .....   | 10         |
| 1.2     The 'AI evidence pathway': collaboration for evidence.....                              | 20         |
| 1.3     Our motivation for the AI evidence pathway .....  | 25         |
| <b>2     Methodological approach for developing this ontology .....</b>                         | <b>31</b>  |
| 2.1     Premises.....   | 31         |
| 2.2     Derivation of nine ethical principles plus trust as a desideratum.....                  | 32         |
| 2.3     Fundamental concepts associated with ethical principles and translational concepts..... | 40         |
| <b>3     Ontology A: Ethical principles and translational concepts.....</b>                     | <b>42</b>  |
| A.0     Trust and trustworthiness .....   | 42         |
| A.1     Beneficence .....   | 53         |
| A.2     Non-maleficence .....   | 66         |
| A.3     Dignity, freedom and autonomy .....   | 84         |
| A.4     Privacy protection.....   | 98         |
| A.5     Transparency .....  | 108        |
| A.6     Responsibility .....  | 138        |
| A.7     Fairness.....   | 151        |
| A.8     Solidarity .....  | 162        |
| A.9     Sustainability.....   | 171        |
| <b>4     Ontology B: Fundamental concepts .....</b>   | <b>174</b> |
| B.1     AI and AI systems.....  | 174        |
| B.2     Ethics, AI ethics, governance, management .....   | 187        |
| B.3     AI actors and communities.....  | 214        |
| B.4     Agency, autonomy and automation.....  | 222        |
| B.5     Bias, heuristics, drift & shift.....  | 231        |
| B.6     AI Evidence pathway, AI life cycle and AI value chain .....                             | 241        |
| B.7     Data.....   | 252        |
| B.8     Algorithm, model, algorithm-to-model transition.....                                    | 269        |
| B.9     Relevance .....   | 309        |
| B.10     Verification, validation, evaluation .....   | 312        |
| B.11     Clinical concepts .....  | 326        |
| B.12     Use of AI systems in health and healthcare .....                                       | 340        |
| <b>5     Conclusion .....</b>   | <b>350</b> |
| <b>References .....</b>   | <b>351</b> |
| <b>List of abbreviations and definitions .....</b>  | <b>402</b> |
| <b>List of boxes.....</b>   | <b>404</b> |
| <b>List of figures.....</b>   | <b>405</b> |
| <b>List of tables.....</b>  | <b>409</b> |
| <b>Annexes.....</b>   | <b>410</b> |
| Annex 1 – Tools for clinical studies and evaluation of AI in healthcare.....                    | 410        |
| Annex 2 - Detailed table of contents.....   | 417        |
| Annex 3 - Alphabetical list of entities of this ontology .....                                  | 425        |

## Abstract

Health, inherently rich in multi-modal data, could profit significantly from artificial intelligence (AI)<sup>1</sup>. Yet, adoption of AI in health remains challenging due to three key issues: (1) The “trust barrier”: while a plethora of documents based on AI (ethical) principles are available, there remains a significant interpretation gap between high-level desiderata and detailed actionable concepts<sup>2</sup>. This hampers determination of both type and level of evidence that would render AI tools sufficiently trustworthy for adoption and integration into use contexts and environments. This is further complicated by the heterogeneous landscape of principles used by various organisations - despite robust evidence on a convergence towards ca 10 principles<sup>3</sup>. (2) The “complexity barrier”: health is complex in terms of life cycle and value chains, involving specialised communities that need to develop and translate AI governance into pragmatic approaches that integrate up- and downstream life cycle stages in terms of evidence requirements<sup>4</sup>. This requires networked thinking, forward-looking planning and bridging of disciplines and domains<sup>5</sup>. However, out-of-domain literacy is typically limited, impeding effective collaboration for trustworthy AI. (3) The “technical barrier”: interoperability and infrastructure needs that may collide with the underfunding of health systems.

To tackle these issues, we propose an ‘AI evidence pathway for health’ aimed at collaboration for evidence on trustworthy AI. The present ontology is its cornerstone. It lays out a pathway for evidence identification, using 10 consensus ethical principles<sup>6</sup> which are unfolded into 42 high-level ‘translational concepts’ that branch into further 110 lower-level concepts (part A of the ontology). The translational concepts connect to 12 clusters of 179 fundamental socio-ethical, scientific, technical, and clinical concepts<sup>7</sup> relevant for AI design, development, evaluation, use and monitoring (part B). Relationships between individual concepts are indicated throughout. The ontology defines user communities for AI innovation in health and outlines a comprehensive life cycle and value chain framework. We introduce the concept of “algorithm-to-model transition” to capture all decisions that may impact on benefits and risks of a model – throughout the life cycle and across value chains. The ontology embraces the benefit-risk ratio concept<sup>8</sup>, emphasising the need for robust real-world evidence on possible benefits of AI tools. The concept descriptions are enriched by a total of ca. 900 publication references. The ontology provides an innovative and comprehensive knowledge resource to support the bridging of relevant actor communities and foster collaboration in view of ‘operationalising’ trustworthy AI in health.

---

<sup>1</sup> OECD (2024) AI in health. Huge potential, huge risks.

<sup>2</sup> Widjaja JT (2024)

<sup>3</sup> Jobin A et al. (2019)

<sup>4</sup> Evidence on absence of risks is important but insufficient. Health products (including AI-enabled ones) rely typically on a benefit-risk ratio evaluation. Thus, robust evidence on added value and real-world benefits is critical and needs to be planned for at early design stages so that it is available for downstream stages such as health technology assessment, deciding on reimbursement and, thus, to an extent, return on investment.

<sup>5</sup> OECD (2024) Collective action for responsible AI in health.

<sup>6</sup> The principles are based on the general consensus identified by Jobin et al., (2019) and are: 1) Trust & trustworthiness. 2) Beneficence. 3) Non-maleficence. 4) Dignity, freedom & autonomy. 5) Privacy protection. 6) Transparency. 7) Responsibility. 8) Fairness. 9) Solidarity. 10) Sustainability

<sup>7</sup> 1) The 12 clusters are: 1) AI & AI systems, 2) ethics, AI ethics, governance & management, 3) actors & communities, 4) agency & autonomy, 5) bias, heuristics & drift / shift, 6) AI evidence pathway, life cycle & value chain, 7) data, 8) algorithm, model & algorithm-to-model transition, 9) relevance, 10) verification, validation & evaluation, 11) clinical concepts and 12) use of AI in health and healthcare

<sup>8</sup> See for instance Annex 1, chapter 1, point 2 of the EU's medical devices Regulation (EU, 2017a).

## **How to navigate this document**

For ease of navigating the pdf version of this ontology, we suggest

- To use **Annex 2, which provides a detailed table of contents**
- To activate the **bookmark panel** on the left of the Adobe Acrobat Reader window. This will allow constant access to the organisation of the entries of the ontology and, by clicking on a bookmark link, allow the reader to jump to a specific point of interest.
- In addition, **Annex 3** lists all 331 entries of the ontology in alphabetical order and links the listed terms to their concept description (control + left click).

## Foreword

The myth of the golem, the man-made clay creature is a compelling metaphor for the tension between artificial intelligence (AI) and human oversight. It illustrates the challenge of reaping AI's benefits while controlling its risks. Like AI, the golem can be a powerful helper, but may turn into a threat if not properly supervised. There are intriguing metaphoric parallels, for instance between the *incantation* of the golem and the development of AI *algorithmic code* or the need to give the golem precise orders and the importance of clear prompts when interrogating large language models<sup>9,10</sup>.

Health and healthcare perhaps represent a litmus test of our capacity to tame the “golem” of AI so that its formidable powers remain in the service of all. Health is inherently rich in multimodal data. The unique capacity of ‘traditional’ and generative AI to analyse data, to integrate information and to generate predictions or novel ‘content’ (i.e. data predictions with contextual meaning for humans) could revolutionise healthcare, health system management, public health surveillance and health research. It could address pressing global socioeconomic challenges of demographic change, a rise in non-communicable chronic diseases and workforce shortenings.

However, whether these positive impacts are becoming reality is not yet certain. It will depend on whether the multidisciplinary community of AI developers, clinicians, researchers and regulators and policy makers will be able to balance AI's many potential benefits versus its ethical pitfalls and considerable risks. These include data and concept bias, propagation of health inequities, de-skilling of clinicians, and deterioration of the human dimension of care, automation bias and complacency, use of unnecessarily complex models and subsequent lack of interpretability. A lack of understanding of AI systems, their applications and limitations can have adverse impacts on their safe use. In general discussion, AI models are often endowed with anthropomorphic adjectives or personification - while even the most sophisticated large language model remains a stochastic automaton of computer code. This creates irrational reactions: unbridled trust as well as deep fears or the feeling of being powerlessness vis-à-vis a ‘superior’ machine. It is critical to confront AI in a rational way, working out guardrails within algorithms but also in respect to training and responsible use of AI systems in real-world use.

This tension between benefits and risks is only to some extent a regulatory or legislative question of product harmonisation but depends also on a mind-set of responsible use<sup>11</sup> within healthcare workflows. This asks for close collaboration between various communities, requiring knowledge of key concepts along the evidence pathway from development to responsible use, knowledge that goes beyond ‘silos’ of domains. This ontology aims at providing a community-bridging tool to facilitate collaboration for evidence to support safe and responsible AI in health.

---

<sup>9</sup> “While we ... create new ... golems, the awful possibility is lurking that they may develop a volition of their own, become spiteful, treacherous, mad golems.” From the article “The golem is a myth for our time” by I.B. Singer that appeared in the New York Times in 1984. Online (paywall): <https://www.nytimes.com/1984/08/12/theater/the-golem-is-a-myth-for-our-time.html>

<sup>10</sup> For an exploration of the digital-metaphoric aspects of the golem myth, see a discussion between Joshua Cohen and Caspar Battegay. Online: <https://www.jmberlin.de/en/joshua-cohen-you-want-a-golem>

<sup>11</sup> OECD (2024) collective action for responsible AI in health. [https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/01/collective-action-for-responsible-ai-in-health\\_9a65136f/f2050177-en.pdf](https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/01/collective-action-for-responsible-ai-in-health_9a65136f/f2050177-en.pdf)

## **Executive summary**

Artificial intelligence (AI) technology represents an enormous opportunity to address a wide array of societal challenges ([OECD, 2024c](#)). Yet, AI requires effective governance frameworks to ensure a positive relationship between benefits and risks. The EU has embraced a comprehensive strategy for AI<sup>12</sup> and, importantly, issued – on a global scale – the first product harmonisation legislation on AI-based products, aiming to foster safe and ethical development and use of AI-enabled products within the EU's single market ([EU, 2024a](#)).

At the root of AI governance initiatives and legislation such as the EU's AI Act or Korea's AI basic Act is the broad concept of '*trustworthy AI*' (e.g. [EU HLEG, 2019](#); [OECD, 2019a](#)). However, considerable effort will be required to 'operationalise' trustworthy AI – through both legislative implementation and 'soft law'. Pragmatic and accessible frameworks, guidance, actionable approaches and, where useful, standards are needed to ensure responsible implementation and use of AI in its real-world settings. After all, not all AI risks are due to intrinsic properties of the technology. Significant risks stem from how AI is used in practice.

Operationalising trustworthy AI in health is perhaps a particular challenge. Health and healthcare are characterised by a highly interdisciplinary community of actors and a complex value chain, resulting in distributed responsibilities that require close collaboration to ensure generation of appropriate evidence on safety, traceability and to clarify accountability in case of problems. Most importantly, health is an area where a lot is at stake – the patient's wellbeing and life. AI applications carry significant risks and are met, understandably, with concern and mistrust. Key issues include the opacity of AI algorithms and the fact that it is the first technology that directly challenges human agency<sup>13</sup>. Potential AI risks in health applications are numerous and include for instance equity and bias, the impact of AI on the doctor-patient relationship (including automation bias and complacency), conceptual and scientific shifts that cause performance drift or, importantly, paucity of sufficient clinical evidence relating to real-world use contexts and environments.

Considering that ethics is about *agency*, i.e. how and why we should act in a certain manner, it is only logical that AI risks are strongly related to *ethical* concepts and human rights, captured as ethical or value-based "principles". The more the AI application is feeding into decisions with far-reaching consequences, e.g. therapeutic choices, complex surgical interventions, health promotion pathways or public health actions, the higher the risk. Thus, finding the balance between benefits and risks of AI requires ethics and ethics-based AI governance. Trust is, as for all technologies, a precondition of AI adoption in real-world settings. Trust requires evidence to render AI solutions trustworthy. Thus, emerging AI governance frameworks will need to focus not only on processes, but on required evidence needs, i.e. on the '*what*', not only the '*how*'. While there is growing international consensus around a broad policy and governance concept of "trustworthy AI", this needs to be put into practice in order to enable the generation and making

---

<sup>12</sup> See European Commission website: European approach to artificial intelligence. Online: <https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence>

<sup>13</sup> The mechanisation of manual labour which started during the industrial revolution had ethical implications on human dignity and jobs. AI has much broader ethical implications due to its capacity to rival human intelligent agency.

available of relevant evidence<sup>14</sup> that satisfies sector or application-specific needs. But what evidence is needed to ‘operationalise’ ethical and hence trustworthy AI?

This question turns out to be rather challenging for essentially two reasons: an inconsistent landscape of principles as well as an interpretation gap between principles and more practical concepts. There has been a plethora of documents by private and public organisations on the nascent concept of AI governance, predominantly based on ethical or value-based “principles”. The OECD dashboard<sup>15</sup> currently counts about 1800 policy-related documents. As pointed out by several authors before<sup>16</sup>, ethical principles *alone* do not ensure ethical safe, robust and equitable AI. Similarly, (life cycle) *processes* do not guarantee the generation of evidence that *matters*, i.e. that is informative and relevant for a specific use context and environment. While there is convergence towards common principles amongst documents, various documents by different (international) organisations or consortia, present these principles in a variety of ways, amalgamate various ethical principles under one new principle or blend ethical principles with technical requirements. This has led to a highly inconsistent landscape of AI ethical principles. As a consequence, AI practitioners wanting to elaborate AI governance approaches and pathways do not have a uniform starting point from which to translate ethics into actionable concepts that can be aligned in dedicated processes. This may complicate international cooperation, harmonisation and recognition of ethical AI approaches. It has therefore not only ethical but also implications for international regulatory science, the emerging concept of *AI governance* as well as trade, competition and competitiveness. This lack of consistency concerning the formulation of ethical or value-based principles further complicates the current ‘interpretation gap’ (Widjaja, 2024) between fundamental ethical principles and practice-oriented, actionable concepts. Although there is demonstrable convergence towards a set of ethical principles (e.g. Jobin et al., 2019), this convergence may hide conceptual and normative disagreements (Mittelstadt, 2019) – since principles are not sufficiently detailed. Clarifying what principles precisely entail will help elucidating potential discord and open avenues for international convergence between actor communities. Since trustworthy AI is based on ethical principles, these need to be ‘unfolded’ into sufficiently detailed scientific, technical and clinical concepts. Once the relevant practice-oriented concepts are clear, the evidence needed to satisfy the desideratum of trustworthiness can be generated and evaluated. Beyond the ‘interpretation gap’, the various communities needing to collaborate for responsible AI in health, require a common understanding of key terms and concepts<sup>17</sup>: data scientists may be unfamiliar with clinical evidence needs, while clinicians will require a basic understanding of AI fundamentals (Gazquez-Garcia et al., 2025).

To tackle these challenges, we propose the ‘*AI evidence pathway for health*’, a conceptual framework for collaborative identification of evidence needs and generation of evidence across the life cycle and value chain. The present ontology is a cornerstone of this pathway. The ontology is designed to bridge the interpretation gap by providing a *conceptual ontology* of translational concepts branching off from a set of *ten granular consensus ethical principles*.

---

<sup>14</sup> OECD (2024) AI in health. Huge potential, huge risks. Online: [https://www.oecd.org/en/publications/ai-in-health\\_2f709270-en.html](https://www.oecd.org/en/publications/ai-in-health_2f709270-en.html)

<sup>15</sup> OECD AI policy observatory – policy instruments. Online: <https://oecd.ai/en/dashboards/overview/policy>

<sup>16</sup> See for instance: Mittelstadt BD (2019) or Morley et al., (2020a)

<sup>17</sup> The guidance on trustworthy AI by WHO (WHO, 2021a,b) and the recent FUTURE-AI consensus guideline for trustworthy AI (FUTURE-AI, 2025) are important steps in that direction. However FUTURE-AI is not aligned with WHO’s or other principles. Moreover, both guidelines use composite ‘principles’ that mix ethical principles with technical requirements. Both provide health-specific recommendations for AI development, deployment and, to some extent, assessment.

These are based mainly on the 11 principles identified by [Jobin and co-authors \(2019\)](#) in their important review of AI ethics guidelines. The proposed 10 principles of this ontology draw also on the work of Morley et al, [\(2020a\)<sup>18</sup>](#), Hagendorff [\(2020\)](#) (in particular in regard to addressing existing gaps), Fjeld et al., [\(2020\)](#), Ryan & Stahl [\(2021\)](#) and Kluge Corrêa et al. [\(2023\)](#). The review by Jobin et al, [\(2019\)](#) included already both the EU high-level expert group ethics guidelines [\(EU HLEG, 2019\)](#) as well as the OECD AI principles [\(OECD, 2019a\)](#) – both highly influential guidance on ethical or value-based AI. Given that we derived the ten granular principles from evidence-based analyses of public AI ethics documents, it is not surprising that these can be fairly easily mapped to the differently presented principles of various guidance documents. This includes AI ethics documents published later than the analyses by Jobin [\(2019\)](#), Hagendorff [\(2020\)](#) and Ryan & Stahl [\(2021\)<sup>19</sup>](#), e.g. the NIST risk assessment framework [\(NIST, 2023\)](#), the playbook of the US department of health and human services [\(2021\)](#) or, recently, the Council of Europe Framework Convention on AI [\(Council of Europe, 2024\)<sup>20</sup>](#).

Against this background, we emphasize that this ontology is not proposing yet another set of ethical principles: we rather use, as a point of departure, convergent singular and granular ethical principles that are common to and encapsulated in various published ethical principles of public organisation, namely the EU's high-level expert group's ethics guidance for trustworthy AI. Using this 'lingua franca' allows to subsequently expound the principles further, by defining relevant associated "translational concepts", i.e. concepts that mediate between high-level ethical and fundamental technical, scientific and clinical concepts. Doing this based on a set of principles by *one organisation* would not have been useful for AI practitioners and the AI scientific community due to the difficulty of cross-mapping concepts. Instead, using a common denominator set of principles allows international debate, exchange and further development. Thus, we follow a scientific and pragmatic rather than policy-driven approach of translating ethical principles into practice. By further relating the translational concepts under ethical principles to "fundamental" scientific, technical, clinical, philosophical, ethical and (health)economic concepts, we build a clear ontological system of high-level ethical desiderata to actionable areas of concern, enabling the identification and generation of relevant evidence for trustworthy AI.

The *nine granular principles* used in this ontology (part A) are: 1) beneficence, 2) non-maleficence, 3) dignity, freedom and autonomy, 4) privacy protection, 5) transparency, 6) responsibility, 7) fairness, 8) solidarity and 9) sustainability<sup>21</sup>. These principles are considered and elaborated as *conditions for trust and trustworthiness* – the latter *per se* not an ethical principle but an essential 'currency' of functioning societies [\(Løgstrup KE, 1956\)](#). The corresponding *fundamental concepts* (part B of this ontology) are grouped in 12 clusters: 1) AI and AI systems. 2) Ethics, AI ethics, governance and management. 3) AI actors and communities. 4) Agency, autonomy and automation. 5) Bias, heuristics, drift and shift. 6) AI evidence pathway, AI life cycle and AI value chain. 7) Data. 8) Algorithm, model and algorithm-to-model transition. 9) Verification and validation. 10) Relevance. 11) Clinical and healthcare aspects. 12) Use of AI

---

<sup>18</sup> Morely et al., [\(2020a\)](#) have suggested that consensus may be found on a *global* level. A comprehensive analysis and mapping of AI principles published by public bodies on a global scale is still missing. The SEA platform [\(2024\)](#) provides web-based tools of linking AI principles (LAIP) but has not presented a comprehensive overview. See also Zeng & Huangfu [\(2019\)](#).

<sup>19</sup> This entails 'dissecting' what we call "composite principles", i.e. statements that subsume several individual ethical concepts or technical requirements under one "ethical" or "value-based" principle.

<sup>20</sup> See section A.0, box 2 for more details.

<sup>21</sup> These principles align to a large extent to those proposed by the European Commission high-level expert group (EU HLEG, 2019), see Table 2 in section 2.2.

systems in health and healthcare. All entities or terms of this ontology are also listed alphabetically (Annex 3).

We hope that this ontology will be useful as a community-bridging tool to provide the multidisciplinary actors on AI in health with a transparent and modular presentation of the many important concepts and concept relationships that are relevant for harvesting the considerable potential of AI in health, - from development, evidence generation through evaluation and validation to uptake, implementation and use in real-world settings.

## **Acknowledgements**

This document was conceptualised and developed in the context of the portfolio on “Innovation in Life and Health Sciences” of the European Commission’s Joint Research Centre during the work programme cycle 2023-2024. We thank the portfolio colleagues for feedback on the initial concept of this ontology. We are grateful for valuable technical comments by Valentin Comte on the final text of cluster B.8. We are much indebted to Heidi Olsson for her help with formatting the document.

## **Authors**

Claudius Benedict Griesinger

Vittorio Reina

Dimitrios Panidis

Hubert Chassaigne

## **Author contributions**

Conceptualisation, methodology, ontology architecture: CBG. Derivation of ethical principles and translational concepts: CBG, VR (for data privacy). Literature research: CBG, VR (for privacy, cybersecurity), HC (global legislative and regulatory frameworks, national AI guidance, contributions to literature on liability legislation), DP (contributions to regulatory concepts on medical devices). Drafting of ontology entries: CBG, VR (privacy, responsibility) and DP (parts of clinical and regulatory aspects regarding medical devices). Annex 1 DP and CBG. All authors have read and agreed to the published version of the manuscript.

## **Use of artificial intelligence**

All text in this ontology is ‘AI-free’: it has been drafted without any recourse to AI language models. However, JRC in-house models have been used in a few cases for interrogating few specific terms and concepts of cluster B.8. The texts retrieved from these models were analysed by the authors and used to conduct more targeted literature searches, using common engines and relevant publication databases.

# 1 Introduction

## 1.1 About this ontology

### 1.1.1 The ontology as a foundation of an 'AI evidence pathway for health'

This ontology is a central part of a conceptual framework which we call '**AI evidence pathway for health**' (for details, see section 1.2). The pathway is intended to support the identification of information or evidence required to ensure trustworthy AI solutions and the generation of that evidence - from conception of an AI system over its assessment to its adoption.

**Evidence generation requires** collaboration of highly interdisciplinary communities involved in AI development, deployment, assessment and use - both along and across the life cycle and value chain. However such collaboration is only possible if grounded in a **common understanding of key concepts of ethical, scientific, technical and clinical nature**.

The pathway has five elements, structured in three thematic blocks that address these requirements (**Figure 1**)

#### **Common understanding**

- 1) an **ontology of ethical principles and translational concepts** (part A of the ontology) that connect to concepts of
- 2) an **ontology of fundamental concepts** (part B of the ontology).

**Figure 2** shows the two-layered design of this ontology.

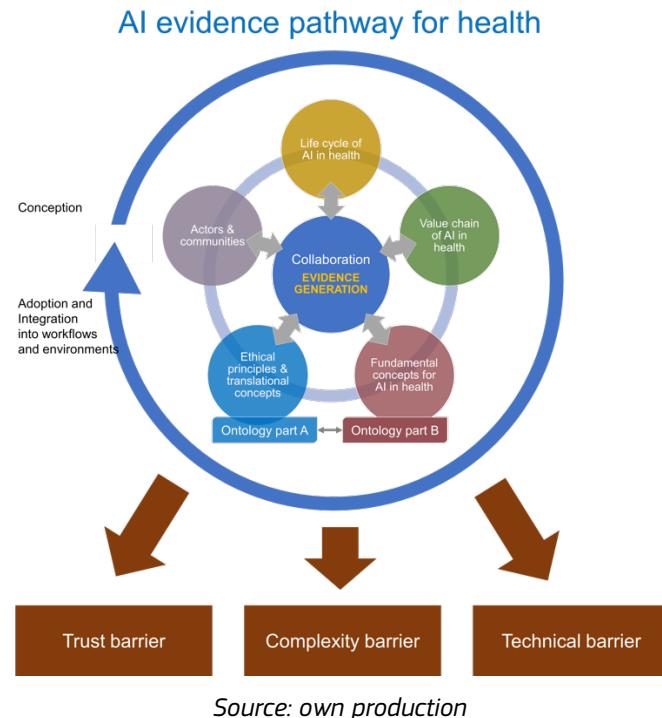
#### **Collaboration**

- 3) **An understanding of relevant actors and communities** that need to collaborate along the life cycle and across the value chain to generate the necessary evidence for trustworthy AI (see B.3 → **AI actors and communities**) and which need to communicate and advocate evidence needs required for specific aspects of the life cycle, e.g. health technology assessment.

#### **Processes and interrelationship**

- 4) The **life cycle of AI in health**, i.e. a widely used image for capturing the complex trajectory of progressing from the conception and design of a solution to its use, monitoring and broader societal evaluation (see B.6 → **life cycle of AI in health**).
- 5) The **value chain of AI**, i.e. enablers (e.g. IT infrastructure, enabling technologies, cyber-security) and assets (data, models) that are required for creating value, e.g. AI systems and associated services (see B.6 → **value chain of AI**).

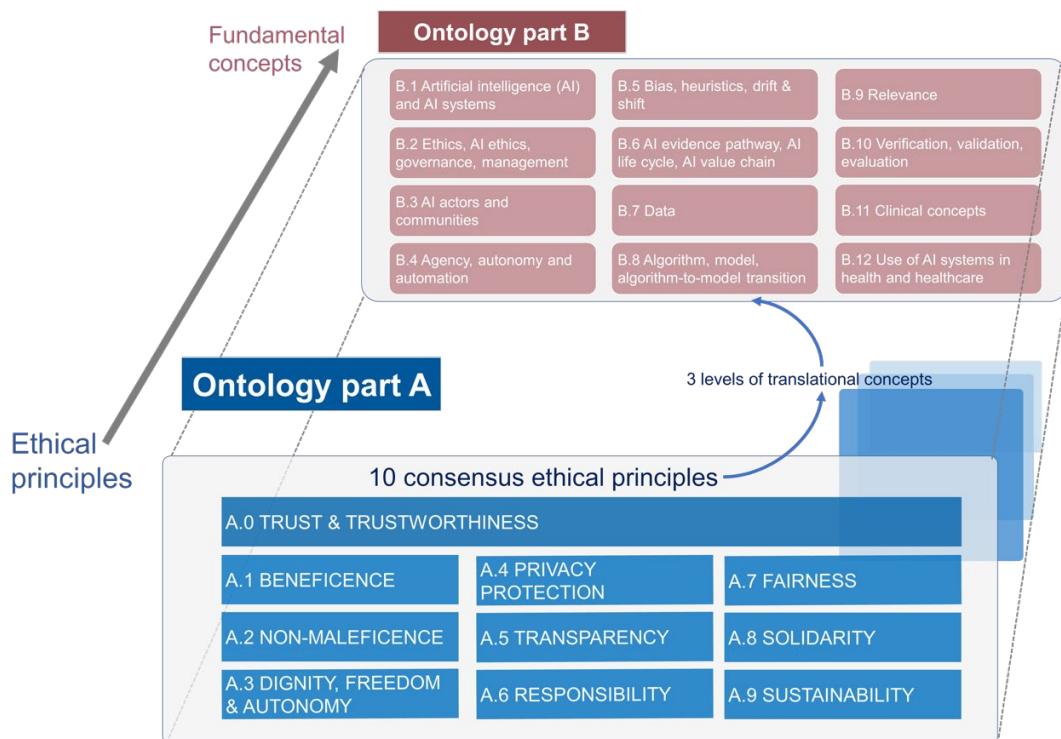
**Figure 1:** The five elements (see Box 1) for enabling effective collaboration and evidence generation along the AI evidence pathway for health.



Source: own production

**Figure 2:** The community-bridging ontology as a fundamental element of the AI evidence pathway for health. The ontology has a two-layer design: ontology A lays out ten ethical principles and their translational concepts (three levels). These connect to ontology B (second layer) which describes fundamental socioethical, clinical and technical concepts, organised in 12 clusters.

### AI evidence pathway for health: community-bridging ontology



Source: own production

### **1.1.2 Application areas of AI in health**

With '**AI in health**' we refer to **four applications areas** based on those outlined by WHO ([WHO, 2021a](#)):

1. healthcare (with a focus on diagnosis and clinical care)
2. health research (e.g. understanding the role of comorbidities for disease progression, research into disease pathways or disease risks) and development of medicinal products ('drugs') and medical devices (whether or not enabled by AI)
3. public health and its surveillance
4. health systems management and planning (including management at individual healthcare settings)

### **1.1.3 About the term 'ontology'**

The term 'ontology' in philosophy means the study of being, or "what there is". Ontologies provide a conceptual organisation of 'entities' and explore their meaning and interrelationship<sup>22</sup>. Ontological entities can be various 'things', e.g. objects, facts, observations, assumptions, perceptions, ideas, mental concepts etc. Ontologies can contribute to forming the required mental categories for understanding, describing and communicating concepts about a given topic.

We use the term 'ontology' in above sense: to explore ethical and technical, scientific and clinical concepts of artificial intelligence in health and their interrelationship with each other. Notably, this ontology should not be mistaken for a machine-readable "ontology" in the secondary adapted meaning of the term, which is now widely used in data science and AI model development<sup>23</sup>.

### **1.1.4 Objective of this ontology**

The ultimate aim of this ontology is to support work towards realising **trustworthy AI** in health. Trustworthy AI is defined via ethical or value-based principles. Based on this, our ontology provides a coherent organisation of **ten consensus ethical principles** and their branching into **lower-level translational concepts** (part A of the ontology). These point to scientific, clinical, ethical, philosophical and technical **fundamental concepts** (part B of the ontology). This closes the current gap between high-level principles and practicable concepts in health.

Trust and trustworthiness of AI depend on **evidence** of these ethical principles and translational concepts. Principles alone are insufficient and may hide divergent views on what they actually entail ([Mittelstadt, 2019](#)). Our ontology thus contributes to the need for translating requirements into technical approaches for trustworthy AI, as called for in the ethics guidelines by the EU's independent high-level expert group ([EU HLEG, 2019](#)<sup>24</sup>). Evidence relating to the ethical principles

---

<sup>22</sup> According to the Oxford Living Dictionary, an ontology is defined as "*a set of concepts and categories in a subject area or domain that shows their properties and the relations between them.*"

<sup>23</sup> N.B. In data science, machine learning and AI, the traditional meaning of the term 'ontology' has been adapted to refer to a precisely defined and constrained model of concepts that relates to a specific real-life phenomenon ([Arnold, 2021](#)). This understanding is rooted in philosophical ontologies. However, such 'ontologies' aim at "*machine-processable semantics of information sources that can be communicated between different agents (software and humans)*", allowing the construction of data domains that "capture" knowledge and which are subsequently manipulated in algorithmic form to permit "*knowledge sharing and reuse*" ([Fensel 2001](#)).

<sup>24</sup> Page 21, section 2.1 Technical methods: "*Requirements for Trustworthy AI should be "translated" into procedures and/or constraints on procedures, which should be anchored in the AI system's architecture.*"

translational concepts is a prerequisite for trust and hence for **the adoption and uptake of AI solutions** (e.g. as a diagnostic tool, augmenting a radiologist's workflow).

The evidence should reflect

- the **application area within health**: healthcare, health research, health systems management and planning, public health ([WHO, 2021a](#)).
- the **fact that health is a high-risk application area**: a lot is at stake. This holds true not just for healthcare but also health research, health systems management and planning as well as public health.
- the particular **risks related to implementing AI systems in clinical workflows** in the case of healthcare applications
- the **specifics of AI in healthcare**, e.g. a uniquely complex value chain and life cycle,
- The **specific AI solution / technology** at hand.

Depending on the AI system's application area and the specific situation at hand, not all of the translational concepts will necessarily be relevant.

In this ontology we address: **a) technological concepts and 'fixes'** related to ethical principles of AI (e.g. → **bias mitigation solutions**); **b) aspects that relate to people and contexts of development, use and deployment** – in particular in situations of *vulnerabilities* and *care relationships* (e.g. impact of AI on patient-physician relationship); **c) distributed responsibilities** and possible resulting risks: a common understanding of key concepts and evidence requirements is crucial in this context.

Above approach is motivated by salient criticisms that AI ethics is too much focused on isolating and “abstracting away” ([Selbst et al., 2019a](#)) complex *ethical problems* through *technological approaches* ([see also Hagendorff \(2020\)](#)). Our approach drew inspiration from a proposal by Aizenberg & van den Hoven ([2020](#)), connecting human rights in AI to design choices of AI.

### 1.1.5 In scope and out of scope aspects of this ontology

#### *In scope*

As outlined above, this ontology is devised as a **scientific concept mapping resource** and fundamental pillar of the → **AI evidence pathway for health**:

- **The ontology is intended for all actors engaged in translating higher-level demands into practical actions**, e.g. when devising relevant processes or guidance. This includes people and organisations developing, deploying and using AI in health, including biomedical scientists, data scientists, modellers, (clinical) researchers, clinicians, medical associations, patient associations (and hence patients), health-technology assessors, ethicists and policy makers and regulators.
- **The ontology has been designed to facilitate collaboration concerning the evidence** required to satisfy the principles of **trustworthy AI in health**. This concerns in particular aspects that are outside the immediate scope of product safety, such as dimensions of use and implementation in health or impacts on patient-physician relationship.

Finally, since AI's benefits and risks are not only determined by the technology itself but to a very large extent by the way we use it. Since AI is constantly evolving, the topic of trustworthy AI in health (and other areas) needs to be regularly revisited.

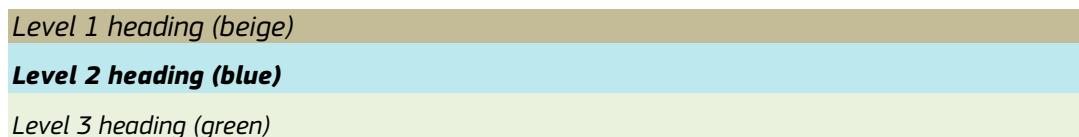
### **Out of scope**

- This ontology is *not a regulatory document*. The primary aim of this ontology is to provide a community-bridging resource of concept descriptions of socioethical, scientific and clinical concepts. As such, it may be potentially useful also for regulatory science.
- This ontology is neither a *terminology* nor a *glossary*. It does not provide an (alphabetical) list of *definitions*. We believe that 'concept descriptions' at this point in time are more useful, given the fluidity and rapidity of change in the AI field. Terminologies typically provide rather short definitions that may be difficult to grasp for the non-expert. They do usually neither provide references to the relevant nor explain term relationships.
- The ontology should also not be understood as a 'taxonomy'. While we provide a tree- and hence taxonomy-like organisation of translational concepts for each ethical principle, this organisation is merely a tool for 'unfolding' and visualising aspects that are encapsulated in ethical principles<sup>25</sup>. Other approaches are conceivable.
- We acknowledge that there is inevitably some degree of overlap between some ethical principles and also between translational concepts. This is a valid criticism of any → **principiplism-based approach** (Clouser & Gert, 1990). However, we hope to make this fact traceable for the reader by visualising such relations in the charts showing the tree-like organisation of ethical principles and translational concepts.

### **1.1.6 Structure of this ontology: two interlinked parts**

The ontology contains a total of **331 concept entries**: part A = **152 entries**, part B = **179 entries**.

- **Part A** deals with an ontology of **ten granular<sup>26</sup> ethical principles and their translational concepts**. These are connected and point to **fundamental concepts** outlined in **part B**. The nine principles are *conditions* for the overarching desideratum of trust and trustworthiness. The translational concepts of part A, are organized in three levels, recognisable by their formatting:



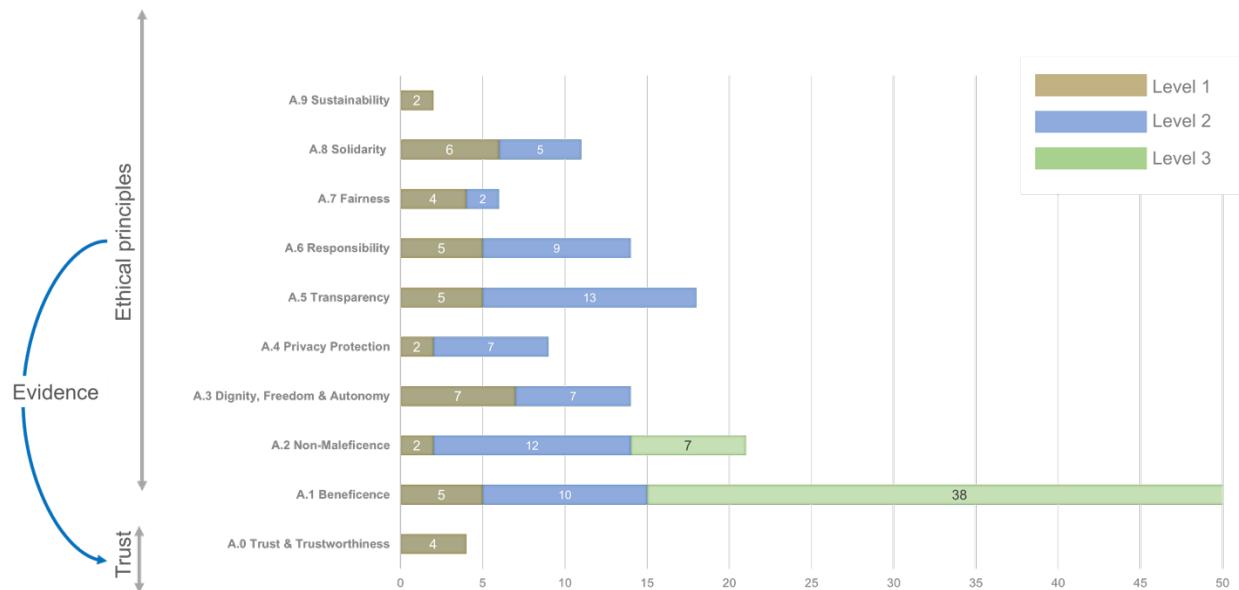
<sup>25</sup> We stress here that other ways of organizing ethical principles and translational concepts are conceivable. For instance, we group 'traceability' as a level 1 translational concept under 'transparency'. From an organizational point of view of quality management, it could however also be placed under the ethical principle of 'responsibility'.

<sup>26</sup> We use 'granular' as an opposite of 'composite', i.e. ethical or AI principles that are composed of more than one principle or desideratum. Thus, for each of the nine principles, one ethical demand can be formulated. This granularity facilitates the translation of demands into actionable concepts.

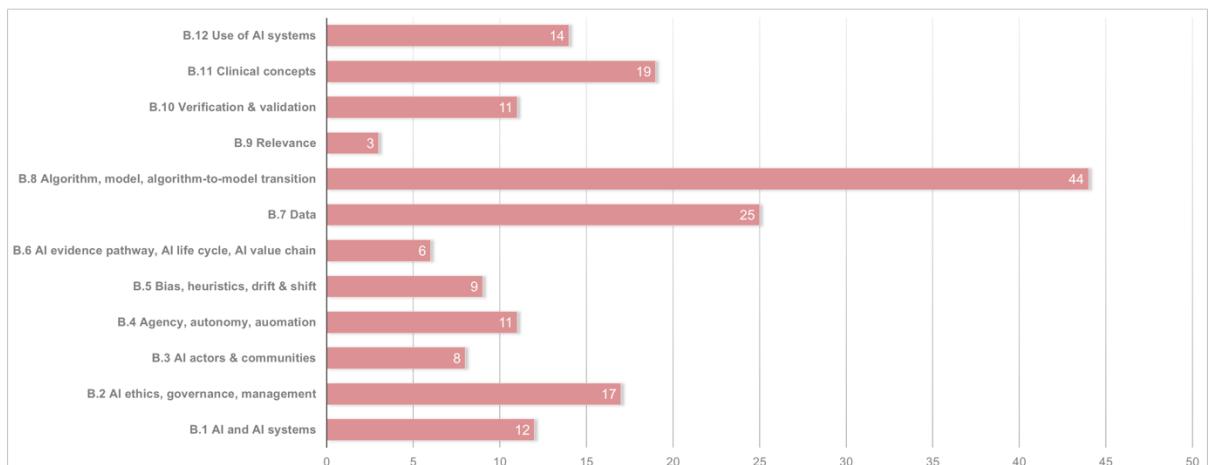
- **Part B** deals with an ontology of fundamental concepts, structured in 12 thematic clusters. These cover ethical, scientific, technical, clinical topics. These are, where necessary, underpinned by relevant policy topics. Foundational international agreements (e.g. Council of Europe, European Union, and United Nations) are briefly explained where necessary.
- Part B should not be understood as a *complete* terminology of AI or → machine learning (which currently does not exist). Instead, its focus is on facilitating the translation of ethical principles into practical concepts in health. We alert to four other existing or emerging terminologies or glossaries for machine learning and AI. Three of these are for AI in general, i.e. not tailored to health; one aimed at AI and digital applications in the health domain. None of these relate the concepts to ethical principles.
  - The Google machine learning glossary ([Google](#)) is widely used by data scientists and AI developers. It is not focused on specific application areas and does not indicate scientific or technical references.
  - The International Organisation for Standardization (ISO) has published a set of definitions and concept descriptions ([ISO, 2022](#)). It is not focused on specific application areas. The references indicated concern mostly other existing ISO standards. Unlike many ISO documents, it is freely available.
  - The EU-U.S. “Terminology and Taxonomy for AI” contains 65 general AI terms in its first iteration. It lists sources used for term definitions ([European Commission, 2023](#)).
  - The FDA has published in 2024 a first version of a “Digital Health and Artificial Intelligence Glossary” which is intended to be an educational resource ([US FDA, 2024a](#)).

**Figure 3: (a)** Ontology A contains 152 translational concepts relating to the 9 ethical principles plus 'trust'. **(b)** Ontology B contains 179 concepts, structured in 12 clusters. In total, the ontology contains 331 concepts.

### a      Ontology part A    Ethical principles and translational concepts



### b      Ontology part B    Fundamental concepts



Source: own production

### 1.1.7 Term layout, formatting and references

The concept descriptions are our own elaboration, relevant references are indicated. For a few entities, existing descriptions or definitions have been used (e.g. thematic cluster clinical concepts B.11 “clinical and healthcare aspects” contains definitions of the International Medical Devices Regulators Forum). Direct citations are indicated by quotation marks and printed in italics.

For top-level entries of this ontology (i.e. all ten ethical principles and all terms in ontology B), we provide the following information:

- A **concept description**: A description of the term’s meaning, highlighting the term’s specific significance in regard to the health context.
- An **explanatory note**, providing further background or context.
- Indications to **literature and further reading**: we provide references to publications which we used for the concept descriptions, and we point to relevant biomedical, clinical publications in the health domain as well as documents on law, policy, economics and philosophical foundations, AI and data science and technology, where needed.
- **Related terms or synonyms** indicate interrelationship with other terms.

Concepts and their relationships are highlighted as follows:

- Throughout the ontology, terms covered by a dedicated entry in the ontology appear in red font, preceded by an arrow, e.g. “→ Ensuring the means for free and informed consent” or “→ Principlism.”
- Terms indicating the ten principles are in capital letters, e.g. “→ DIGNITY, FREEDOM AND AUTONOMY” or “→ NON-MALEFICENCE”.

Lower-level translational concepts under a given ethical principle are presented without above structure.

For most entries we provide references to relevant publications, listed in the reference section. For most references, links for online access are provided.

The ethical ontology (part A) points extensively to fundamental technical, scientific, clinical and ethical concepts provided in ontology part B. However, also entries in part B are contextualized by references to ethical principles in part A. We hope that this modular approach facilitates networked thinking required for AI in health.

For instance, in separate sections on the ethical principles of → FAIRNESS and → DIGNITY, FREEDOM & AUTONOMY, we refer to various forms of → bias, which are described in more detail in one single section on → bias, heuristics, drift & shift (part B).

Although tailored to AI practitioners in health and healthcare, the present ontology may be useful for other AI application areas.

## 1.1.8 Practical applications of the ontology

### a) A foundation of the 'AI evidence pathway'

First, this ontology is designed as a **fundamental part of a broader conceptual framework, the → AI evidence pathway for health** (see section II) by providing common understanding of terms and their relationship when designing and executing processes along the → **life cycle of AI in health** and across the → **value chain of AI** related to design, development, production, deployment, evaluation & assessment, decommissioning. It might support ongoing developments towards general and more application or sector-specific approaches towards 'AI governance'. The need for governance of AI-empowered healthcare has been highlighted ([Reddy et al., 2020b](#); [Čartolovni et al., 2022](#); [Reddy, 2024](#)).

### b) An educational tool for bridging multidisciplinary communities

Second, this ontology might be useful as a tool for **education and to bridge heterogeneous and multidisciplinary communities that need to collaborate to realise a responsible use of AI in health** ([OECD, 2024d](#)). Realising ethical and trustworthy AI through collaboration requires a common conceptual understanding. The need for common terminologies (e.g. [Griesinger et al., 2022](#)) and for education (e.g. [Council of Europe, 2022](#)) has been emphasized in various documents.

### c) A conceptual framework to support benefit-risk considerations of AI in healthcare

Third, this ontology should be seen in the context of the reality that the **considerable possible benefits of AI in healthcare and other health applications come with potentially considerable risks** ([OECD, 2024c](#)). Most AI ethics guidelines have been focusing on AI's risks. However, sufficiently robust **evidence on benefits** is critical for healthcare applications for at least two reasons:

- First, **acceptability of residual risks** of health products (e.g. medical devices, including AI-enabled ones) is often defined based on an overall **benefit-risk ratio**<sup>27</sup>, generated during → **clinical evaluation** (e.g. EU's medical device Regulation: Annex I, Chapter 1, point 2; [EU, 2017a](#)).

AI introduces many new challenges due to its **complex value chain** and resulting highly **distributed responsibilities** (e.g. data from suppliers, pretrained models, device producers, service providers), complicating risk control. Robust evidence on benefits will help contextualizing risks, which cannot be completely eliminated without negatively affecting benefits.

Importantly, absence of risks does not yet constitute an added value. Only by having reliable evidence on benefits, various communities, notably clinicians and health systems can decide on whether or not these benefits outweigh the risks and decide whether to adopt and rely on AI solutions in clinical workflows, for clinical administration, health system management and planning and public health surveillance and measures.

---

<sup>27</sup> There is no formula or mathematically precise concept of *benefit-risk ratio*. The term should be understood as the integrative weighing of likely or demonstrated benefits versus likely or demonstrated risks.

- Second, in many countries AI tools used in healthcare will, as other health technologies, require **socioeconomic, clinical and ethical assessment** prior to reimbursement decisions within health systems ('health technology assessment', HTA). Such comprehensive assessments require robust clinical evidence and draw *inter alia* on the benefit-risk ratio.

Thus, by including considerations and a set of concepts on benefits under the ethical principle of → **BENEFICENCE**, we suggest a more systematic categorisation and assessment of benefits of AI-enabled health products, supporting the responsible design, deployment and use of AI in various health domains. We stress that these categories are not intended to replace a comprehensive description of benefits of a given product.

The high-level translational concepts of 'beneficence' presented here reflect the evidence progression along the pathway, moving from conceptual '→ added value' and '→ gains' during design phases over '→ potential benefit' during pre-deployment and supported by evidence from optimized conditions in research settings to '→ real-world benefits' supported by robust evidence from real-world → **use environments**. We stress that these are not regulatory terms, but concepts that we propose for scientific considerations, and which are intended to capture the trajectory of available evidence on beneficence over the life cycle.

#### **d) A forward-looking approach to ensure efficient and safe innovation as well as competitiveness**

Finally, this ontology and the concept of the **AI evidence pathway** for health are intended to support a **forward-looking approach during the life cycle**: already during upstream phases of the life cycle relevant evidence requirements for downstream phases should be sufficiently considered in order to navigate efficiently the transition to a usable AI solution that is trusted and hence adopted and relied on. In the context of the AI evidence pathway, we have previously provided an ontology of health impacts and value chain enablers of digital health solutions ([Reina & Griesinger, 2024](#)).

This approach is fundamental for **safe innovation** (Article 13 of the Framework Convention on AI and Human Rights; [Council of Europe, 2024a](#)) and will support **competitiveness of actors and organisations** by avoiding unnecessary delays at downstream stages of the pathway.

#### **1.1.9 How to use this ontology**

This is intended as a resource for AI practitioners engaged in designing, using, assessing trustworthy AI. Ontology entries can be interrogated as needed:

**Start from ethical principles (part A):** Readers may want to familiarise themselves first with the concept descriptions of the ethical principles and the organisation of translational concepts (part A), before following up on the 12 clusters of specific fundamental concepts (part B). The clusters facilitate access to thematic topics.

**Start from fundamental concepts (part B):** Alternatively, the ontology can also be used by interrogating fundamental concepts (part B). However, since individual concepts (e.g. 'data') may relate to many different translational concepts and ethical principles of part A, approaching the ontology in that manner may not allow bridging the interpretation gap between ethical principles and actionable concepts.

## 1.2 The ‘AI evidence pathway’: collaboration for evidence

### 1.2.1 The AI evidence pathway: overview

The ontology is part of a conceptual framework which we call '**AI evidence pathway**' (**Figure 2**).

#### Box 1. What in a nutshell is the ‘AI evidence pathway for health’?

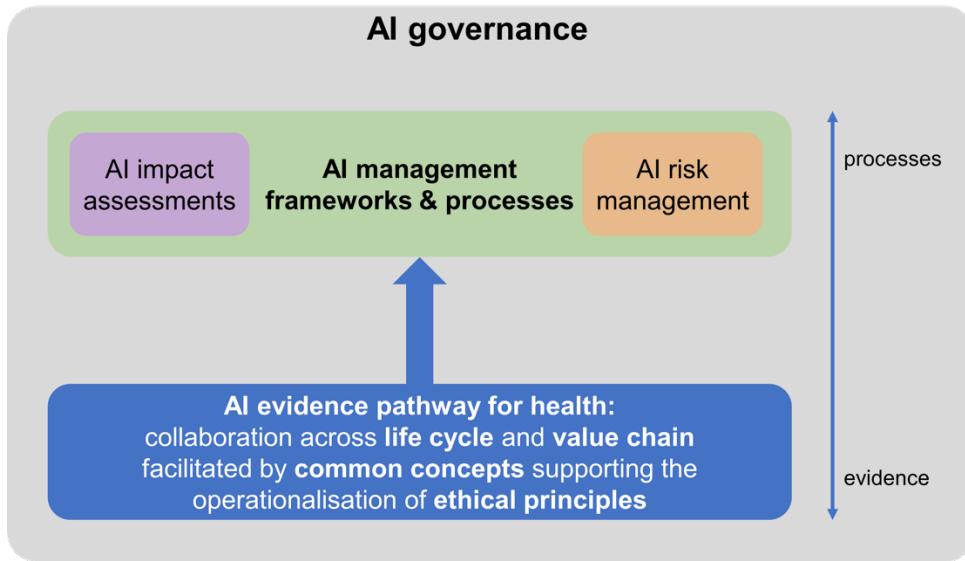
- **Focus on evidence:** Life cycle concepts usually focus on *processes*. The AI evidence pathway focuses on the evidence needs for trustworthy AI along the pathway from conception of an AI solution to its adoption.
- **Collaboration of actors and communities:** To generate this evidence, actors need to collaborate along this pathway in a feedforward and a feedback manner. This will support trust and responsible use of AI in health ([OECD, 2024d](#)).
- **Life cycle of AI in health:** The AI evidence pathway adapts and extends the general life cycle schemes to address health-specific issues, such as real-world implementation in complex environments (e.g. hospitals) the need for health technology assessment and its critical role for innovation.
- **The value chain of AI:** The AI evidence pathway acknowledges the complex value chain of AI in health, which poses risks and uncertainties (e.g. distributed responsibilities). Value chain actors are critical evidence contributors.
- **Bridging the gap between ethical principles and their operationalization:** Collaboration requires a common understanding of key concepts: Trustworthy AI is defined by ethical principles. These are insufficient on their own. They need to be ‘translated’ ([EU HLEG, 2019](#)) into scientific, technical and clinical concepts for operationalising the principles and thus creating trustworthy AI solutions ([OECD, 2024c](#)). Our ontology contributes to this.

We consider the → **AI evidence pathway for health** as a contribution to the overarching concept of → **AI governance** and → **AI management** (**Figure 4**).

Briefly, → **AI governance** refers to the collection of relevant legislative requirements, pragmatic and praxis-oriented approaches (e.g. in specific application domains such as health), principles and frameworks for controlling development and use, monitoring and decommissioning of AI solutions in real-world environments. → **AI management** refers to the practical approaches for realising such control. These may differ between organisations due to the scope of their work and depending on whether they develop AI or use it.

The AI evidence pathway supports the collaboration of various actors and communities involved in AI in health for generating **evidence for trustworthy AI** along the life cycle and across the value chain - from design to adoption. Collaboration is needed for **safe innovation**: it enables feeding real-world experiences and downstream needs to upstream phases of conception and design and is essential for defining evidence needs required to respond to the rapid technological change characteristic of AI.

**Figure 4.** Schematic diagram of the contribution of the AI evidence pathway to AI management by providing a conceptual framework for collaboration of actors to generate evidence for trustworthy AI.

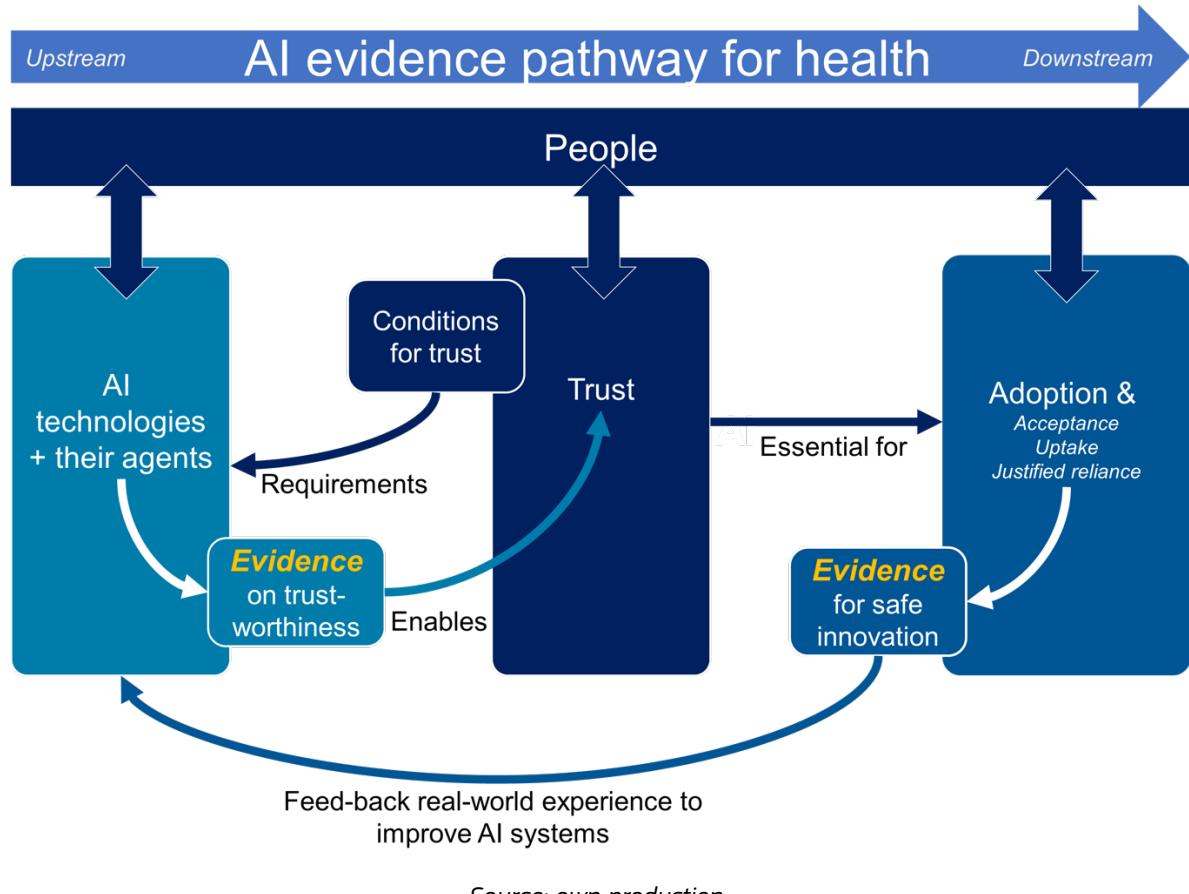


Source: own production

With this pathway concept, we propose a stronger consideration of **evidence generation** through collaboration to support trust in AI (**Figure 5**).

- a) **in a feed-forward manner during early stages of conception and design.** Already at this stage, downstream aspects (e.g. evidence need for health technology assessment) should be considered.
- b) **through the transparent showing of how evidence on trustworthy AI has been realised** and
- c) **through feed-back of experiences and findings based on post-deployment real-world use** into upstream stages of AI conception, design and developments upstream.

**Figure 5.** The AI evidence pathway as the nexus of people, trust, AI technologies and their adoption in health. People develop AI technologies and adopt them. People define conditions of trust and generate evidence supporting trustworthiness. People using AI generate real world evidence for safe innovation.



### 1.2.2 AI evidence pathway: a people-centric and collaborative framework

The **AI evidence pathway is people-centric** (Figure 4): it is **people that are “at the centre of AI”** (OECD, 2024d):

- **People conceive, design, develop, deploy and use AI technologies and decide how to integrate AI solutions in existing workflows, → use contexts and → use environments.** People may be individual actors, organisations which are typically part of communities of practice (e.g. developers, healthcare professionals, users and impacted persons, service providers etc.). Communities typically have their own language, specific ways of working and topics that are of particular interest to them. Effective collaboration across communities requires common understanding of key concepts that cover the evidence needs from conception to adoption ('evidence pathway').
- **People need to have sufficient trust in AI technology to adopt it and be able to justifiably rely on it.** Trust is conditional on supportive evidence that may render both AI actors and their solutions "trustworthy". Conditions of trust include ethical concerns, compliance with legal and regulatory frameworks, agreement with technical, scientific and, where required, clinical requirements.

- **Supportive evidence is generated along the life cycle of AI and requires collaborative efforts.** Data provenance is an example: AI developers and deployers that use data from a supplier require sufficient evidence, e.g. on how data were obtained, how they were processed, to which extents intrinsic biases were checked etc. The same holds for pretrained models, where these are not intrinsically interpretable but require explainability techniques. Collaboration is also needed to specific user requirements, interoperability requirements regarding the use environment and monitoring of performance, safety and undetected bias that may become evident only under real-world use conditions.
- **People generate evidence during real-world use of AI systems** and this evidence should feed back into technology development, to improve intrinsic properties of AI systems and to fine-tune the implementation and integration of AI solutions into workflows and use environments. This feedback supports safe innovation.

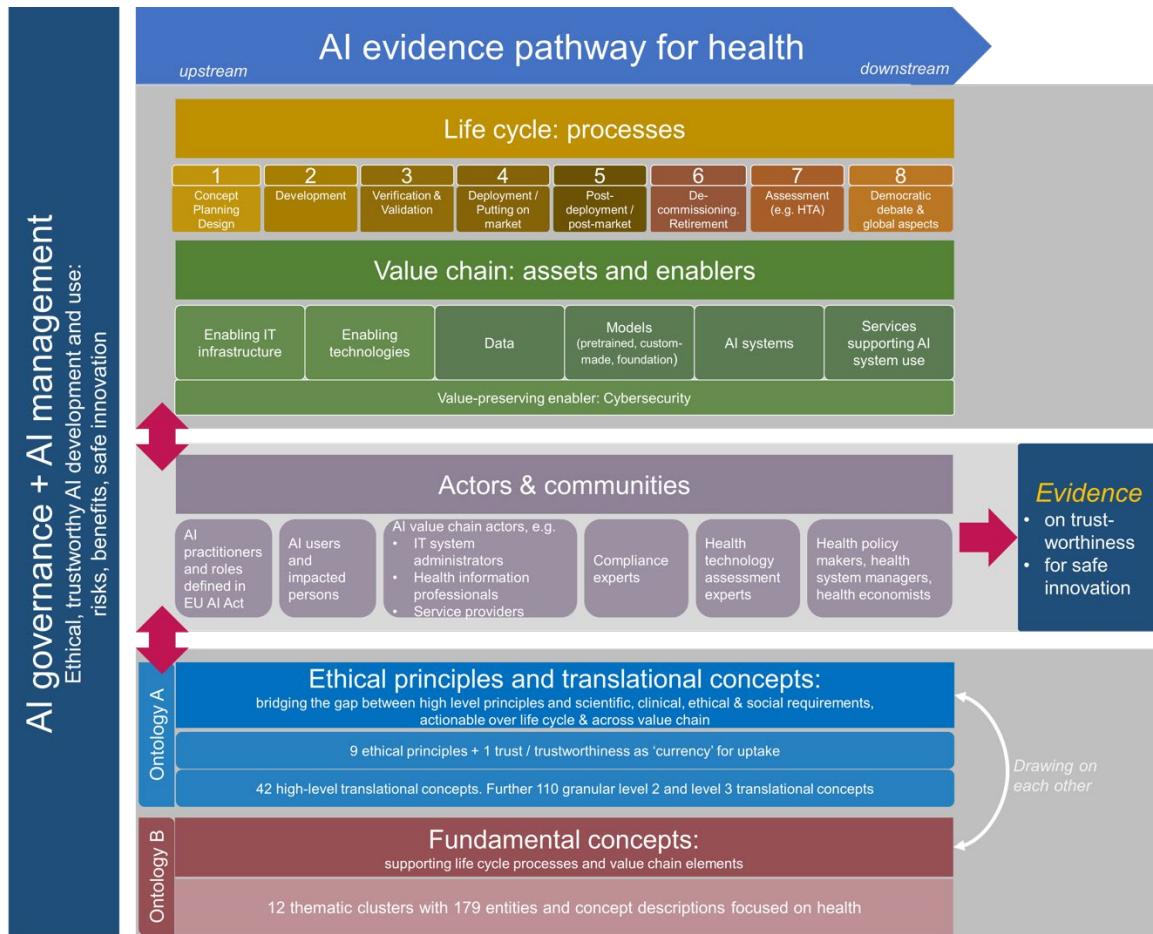
### **1.2.3 The five dimensions of the evidence pathway**

The AI pathway comprises five dimensions of collaboration to generate evidence (see **Figure 6**):

1. **People: actors & communities.** AI in health involves many different experts and communities of practice, both from a life cycle as well as a value chain perspective. Collaboration is key to ensure adoption of safe, secure and performant AI that helps patients, healthcare professionals, health systems and health researcher and that provide true added value.
2. **Life cycle of AI for health.** There is no consensus on a life cycle of AI in health. We propose eight stages for the trajectory from conception, design, and development to deployment and decommissioning. We propose two additional stages further downstream: **health technology assessment** and **open debate**, e.g. about how AI should be used healthcare and health system management. Both stages radiate back into defining needs, limitations and applications of AI in health at upstream stages of conception and design of AI. We therefore consider these two stages also part of the life cycle.
3. **AI value chain for health:** There is no consensus on what constitutes the value chain of AI in health. Yet, to realising an AI solution requires various physical, process and knowledge-based prerequisites which, like the end product, represent value for various stakeholders: AI systems (e.g. for diagnostics, for clinical care and decision making, for analysing big health data) or the services needed to provide these (e.g. telemedicine, wearables, medical training, remote surgery). Actors associated with the AI life cycle (e.g. AI system developers, auditors) need to collaborate with actors of the AI value chain (e.g. providers of pretrained models or data providers) to ensure accurate and complete evidence on key elements used to build an AI system.
4. **Ethical principles (practical AI ethics):** We propose ten ‘granular’ ethical principles based on the analysis of AI ethics guidelines of Jobin et al. (2019). We ‘translate’ these principles into 42 level 1 concepts, with additional two layers of lower-level concepts. All these are organised in an ontological system: relationships to other concepts and ethical principles are indicated.
5. **Fundamental concepts:** The translational concepts relate to about 150 key concepts. These are organised in 12 thematic clusters. The concepts are enriched by references to

clinical, scientific, technical or other types of publications. Some of the concepts represent life cycle milestones and/or processes, e.g. ‘clinical evaluation’, ‘post-deployment monitoring’, ‘decommissioning’ or refer to value chain aspects (e.g. data, model, algorithm, AI system).

**Figure 6.** Schematic depiction of the AI evidence pathway’s five elements required for generating evidence on trust. Agents and communities are at the centre. They generate evidence on trustworthiness and safe innovation by collaborating along the pathway of life cycle stages and value chain elements. The evidence generation is supported by common concepts (ontologies). Evidence from real-world use should feed into new iterations of life cycle processes facilitating safe innovation. Left: AI governance and AI management as overarching concepts, drawing on the AI evidence pathway.



Source: own production

## 1.3 Our motivation for the AI evidence pathway

The approach of formulating ethical ‘principles’ to address moral problems related to technologies dates back at least to the 1970ies with the introduction of ethical principles in biomedicine and life sciences technologies (→ [bioethics](#)). A comparable “→ [principlism-based](#)” approach has been used for AI by various international bodies, countries and private organisations. Those of international public organisations nearly all emphasise **trustworthiness as a key aim**. However, despite the usefulness of these documents for advancing AI ethics, there are challenges for the AI practitioner, which we try to address with the **AI evidence pathway**.

There are, in our opinion, three key challenges:

### 1.3.1 No international consensus on ethical or AI principles

Trustworthy AI is based on ethical considerations: several public bodies have over the last years published relevant guidelines. These all employ a → [principlism-based](#) approach to tackle AI ethics in pursuit of ensuring ‘trustworthy AI’. These guidelines stipulate “ethical” or other (e.g. “value-based”) *principles* for addressing moral concerns associated with AI.

Several analyses have been published comparing ethical guidelines issued by a variety of organisations ([Zheng et al., 2018](#); [Fjeld et al., 2019](#); [Jobin et al., 2019](#); [Hagendorff, 2020](#); [Ryan & Stahl, 2021](#); [Kluge Corrêa et al., 2023](#)). The study by Jobin et al., (2019) is perhaps the most systematic one, demonstrating that there are **11 ethical principles that feature in most ethics documents, indicating a degree of consensus**. The principles are (in order of frequency): *transparency, justice and fairness, non-maleficence, responsibility, privacy, beneficence, freedom and autonomy, trust, dignity, sustainability and solidarity*. Similarly, a review of literature on ethical technology showed that the topics privacy, security, autonomy, justice, human dignity, control of technology and the balance of powers were recurrently used ([Royakkers et al., 2018](#)).

Although these analyses were available already in 2019, guidelines and principles published by public bodies in the following years do not appear to build on these 11 identified consensus principles when presenting and communicating the ethical principles<sup>28</sup>. Thus, the various principles and guidelines from public bodies published over time (see → [TRUST AND TRUSTWORTHINESS](#)) all differ in number, presentation and/or structuring of principles.

This has led to a **heterogeneous landscape of ethical or AI principles**. At the same time, it is fairly clear that – by and large – there is significant agreement between the various documents concerning the key ethical questions raised. Nevertheless, formally there is no single consensus avenue for the AI practitioner in regard to a common conceptual basis of AI ethics principles, which complicates ethical and scientific debate, community-discourse and international exchange and acceptance of evaluations, as well as presentation of research results and, importantly, efforts to **detail the concepts behind these principles** (see section 2 below).

A further complication stems from the fact use of what we all ‘**composite principles**’, i.e. principles that are a blend of several concepts. Examples include: “*Respect for the rule of law, human rights and democratic values, including fairness and privacy*” ([OECD, 2019a, amended 2024](#)) or “*Promote human well-being, human safety and the public interest*” ([WHO, 2021a](#)).

While there are good reasons why public organisations structure (ethical) principles in various ways (e.g. emphasising specific concerns in their area of work), the heterogeneity of principles nevertheless poses practical problems for practitioners, regulators, scientists and stakeholders

---

<sup>28</sup> In 2025 the FUTURE-AI consortium published a guideline for trustworthy and deployable AI in healthcare, starting from a new set of six ‘guiding principles’ which are neither aligned with the ones identified by [Jobin et al., \(2019\)](#), nor with those of [WHO \(2021a\)](#).

wishing to formulate more granular lower level concepts under ethical principles: Which ethical principles from which organisation should be taken as high-level ones for further elaboration? How should one deal with ‘composite principles’, blending more than one ethical demand in one principle?

This situation may pose obstacles for international exchange of information or acceptance of ethical assessments and, on a more basic level, creates confusion and uncertainty of debate in an area that is characterised by rapid and disruptive technological development. Thus, using a common set of consensus principles that truly reflect ethical concerns, would help with the agility of the communities that need to ensure the ethical alignment of AI technology with social, ethical values and legislative requirements, rooted in ethical considerations.

### 1.3.2 Gap between ethical principles and practical concepts

In addition to this **heterogeneous landscape**, it is not always sufficiently clear **what these principles precisely entail**. Although charge questions in relation to the principles have been published (e.g. EU high-level expert group ethics guideline and associated ‘ALTAI’ list; [EU HLEG 2019, 2020](#)), there remains a gap between **high-level ethical principles and practical concepts**, when considering specific needs of a given application area (see section 3).

Several publications have over the year pointed out the need for a ‘translation’ of ethical principles and requirements into more concrete concepts. The European Commission’s independent high-level expert group remarked that “*requirements for trustworthy AI should be “translated” into procedures and/or constraints on procedures, which should be anchored in the AI system’s architecture.*” ([EU HLEG, 2019](#)).

Mittelstadt ([2019](#)) has argued that ethical principles alone are insufficient to ensure ethical AI and that, owing to the significant differences between biomedicine and AI, a principlism-based approach for the latter may run into various problems, e.g. due to “*a lack of common aims and fiduciary duties, professional history and norms, proven methods to translate principles into practice and robust legal and professional accountability mechanisms*” ([Mittelstadt, 2019](#)). Further, significant normative and policy disagreements may hide behind the apparent convergence towards a common set of principles.

Morley et al. ([2020a](#)) have pointed out that principles are insufficient for affecting AI design, suggesting that substantial work was required to “translate” the concepts behind principles into practical approaches, requiring collaborative approaches<sup>29</sup>. Previously authors have observed that efforts are unevenly distributed across areas of concern or the life cycle.

There is a focus on technological fixes, e.g. techniques of explainable AI ([Selbst et al., 2019a](#)). Comparably less efforts are made in regard to vulnerable groups and situations or, overall, societal aspects.

Similarly, Shneiderman ([2020](#)) observed a gap between principles and practice, in particular in regard to human-centred AI. Shneiderman also highlighted the need to move on from high-level statements towards clearer concepts, e.g. for social practices.

Thus, without the necessary further **unfolding of ethical principles**, the ethical demands enshrined in these principles cannot be achieved in practice without a high degree of uncertainty and without risk that various actors follow heterogeneous approaches that are difficult to assess and compare (e.g. for regulators or auditors).

---

<sup>29</sup> “*Bridging together multi-disciplinary researchers into the development process of pro-ethical design tools and methodologies will be essential. A multi-disciplinary approach will help the ethical ML community overcome obstacles concerning social complexity...*” ([Morley et al., 2020a](#))

Furthermore, the **relationship between ethical principles and the emerging concepts of AI management and AI governance** is currently not fully clear: a comprehensive framework has still not emerged. Thus, the need for a consistent and clear framework that bridges AI ethical principles and practical concepts needs to take into consideration also product development processes and hence the concepts of “**life cycle**” and “**value chain**”. Both are complex in the health domain and characterised by a high degree of distributed responsibilities.

### **1.3.3 Most guidelines and ethics documents are not tailored to health**

Most of the ethical principles proposed in relevant guidelines are not tailored to a specific AI application domain (e.g. finance, health, education). This makes sense from a high-level policy perspective of addressing all sorts of AI applications. A notable exception is the WHO’s comprehensive document on ethics and governance for AI for health ([WHO, 2021a](#)).

Challenges that are specific to a given application field are left unaddressed by general AI ethics guidelines, e.g. the connectedness of intelligible AI with the requirement for patients’ informed consent to medical treatments or aspects of the relationship between patient and physician.

It is questionable in our opinion whether the translation of meaningful, translation of ethical principles into practicable and actionable concepts can be achieved for AI in general. Problems that are specific to an application field as health and healthcare will inevitably not be sufficiently considered.

There are obviously significant differences between sectors and application fields regarding specific ethical problems. For instance, problems unique to health include the impact of AI on the patient-physician relationship, the impact of unintelligible AI on informed patient consent or AI complacency that may lead to resorting to automated solutions in the face of ethical conundrums in clinical decision making. These problems are not in the same manner applicable to AI use in education, finance, information retrieval (e.g. search engines) or knowledge extraction through prediction of likely next words, pixels etc. (content synthetisation or ‘generative AI’). Having said that, specific technologies (e.g. generative AI) may have potential applications in health. But this requires still risk-benefit approach taking the specifics of health into account (e.g. [WHO, 2024a](#)).

### **1.3.4 Addressing the three challenges through the AI evidence pathway for health**

Taken together, there is a need to ‘unfold’ the ethical principles into lower-level detailed translational concepts that relate to practical scientific and technical concepts. This translation of principles to practice would support ‘operationalising’ AI ethical principles ([OECD, 2024c](#)).

The current ontology is designed to address this need in health. Further, by embedding this ontology in the framework of the AI evidence pathway, we consider also key elements such as **actors and communities of practice, clinical aspects** relating to AI-enabled healthcare products and, importantly, the **life cycle** and **value chain**.

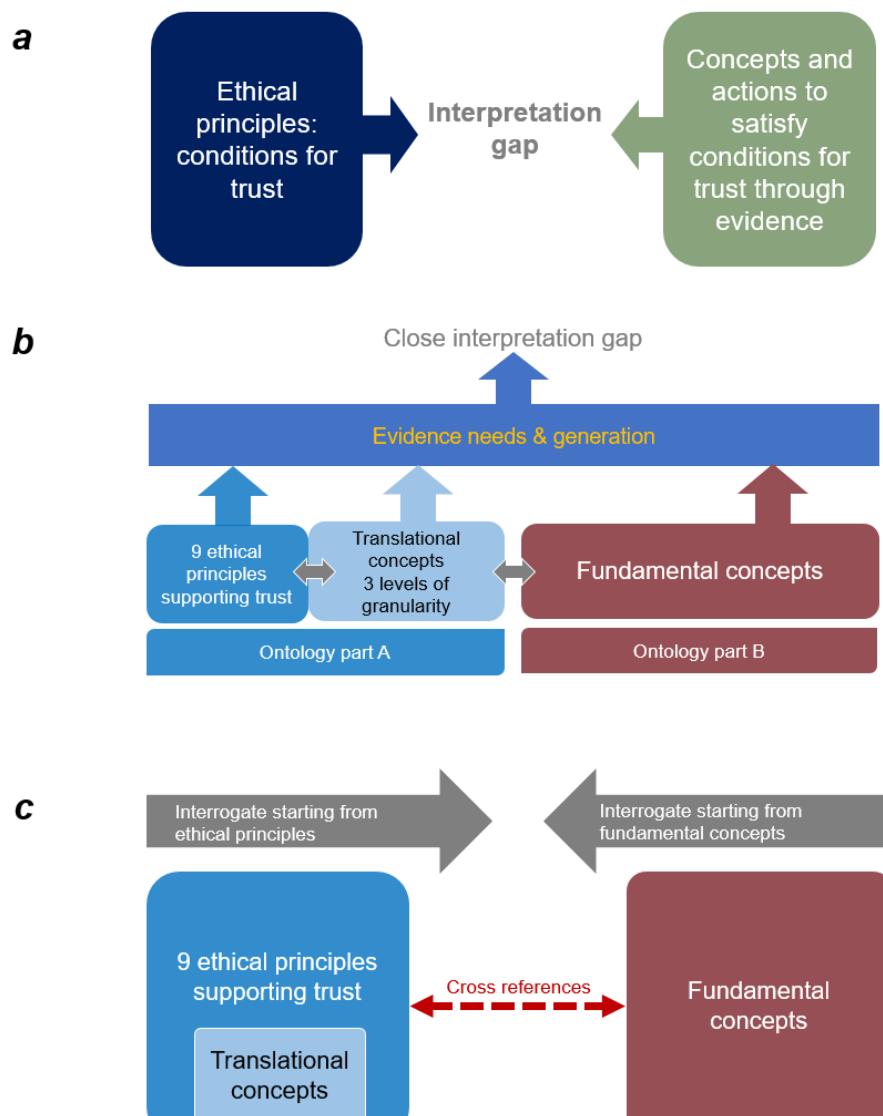
Our proposal relates to previous efforts. For instance, both, the “Interdisciplinary framework to operationalise AI ethics” ([VDE / Bertelsmann Stiftung, 2020](#)) or the “Ethics framework” of DigitalCatapult aimed at providing observables or charge questions for seven ethical principles ([DigitalCatapult, 2023](#)), drawing on the EU high-level expert group’s proposal, including the “ALTAI” check list ([EU HLEG, 2020](#)). [Ryan and Stahl’s \(2021\)](#) overview of AI ethical principles and associated ‘constituent ethical issues’ can be seen as a first step towards an ontological organisation of ethical principles and translational concepts. [Char et al. \(2021\)](#) have provided a valuable outline of key ethical questions in relation to machine learning and AI in health based

on a life cycle approach. Their excellent paper has been an inspiration for our proposal of an AI evidence pathway for health, connecting life cycle, value chain and AI ethics in one framework and putting agents and their collaboration centre stage to generate the required evidence for trust. Morley et al. (2020a) have emphasized the need for moving from the ‘what’ of ethical AI to the ‘how’ of applied ethics, proposing an outline of an applied ethical AI typology.

However, these documents stay on a rather high level and are aimed at different AI applications areas, leaving gaps regarding ethical demands that are specific to application areas, notably health as a high-risk application area.

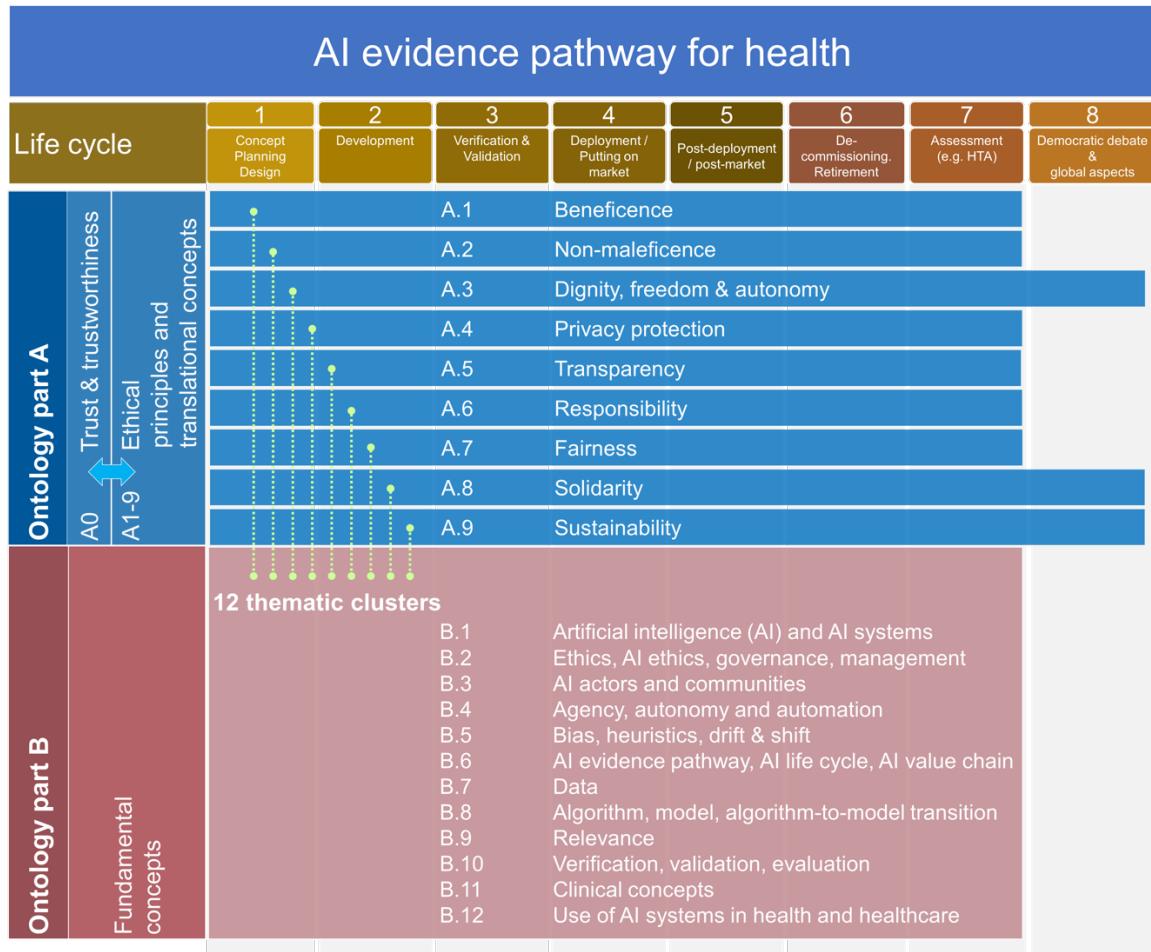
We hope that the AI evidence pathway for health will address these issues by closing or at least narrowing the **current gap between discussions on life cycle, risk management or AI government in health** on one side **and the requirements for ethical AI formulated via stipulation of “ethical principles”** on the other side.

**Figure 7. (a)** The current interpretation gap between ethical principles formulating high-level conditions for trust and concepts to satisfy these conditions through evidence. **(b)** The ontology presented here contributes to closing this interpretation gap in the health area by providing nine granular ethical principles and ‘translational concepts’ that connect to fundamental concepts and aspects including the AI life cycle and the value chain. The ontology is intended to support community-bridging and collaboration to identify evidence needs and generate evidence on trustworthy AI. **(c)** The ontology can be interrogated either from Part A (9 ethical principles and translational concepts) or from part B (fundamental concepts). Both parts are connected via cross references.



*Source: own production*

**Figure 8.** Schematic depiction of the two parts of this ontology: Part A builds on nine ethical principles which are unfolded in translational concepts. These point to fundamental concepts, structured in 12 thematic clusters of part B. Green dotted lines indicate cross connections. Fundamental concepts cover technical, clinical, ethical/philosophical, economic and legal/regulatory topics, covering the life cycle from conception and planning to assessment. Three ethical principles explicitly touch on the stage 'democratic debate & global aspects' (e.g. on the future of AI use in healthcare, sustainability and solidarity issues).



Source: own production

## 2 Methodological approach for developing this ontology

The primary aim of this ontology is to provide a systematic explication (through ‘translational concepts’) of salient ethical topics relating to the use of AI in health. As outlined in 1.3.1, there is no formal consensus of ethical principles from which to start. However, various reviews have established a convergence around a set of ethical principles, which underpin these guidelines. Jobin and colleagues have perhaps tackled this issue in the most systematic manner, identifying 11 key ethical principles that are most frequently named in 84 ethics documents by public and private organisations ([Jobin et al., 2019](#)). This analysis included influential documents by public bodies, including the EU independent high level expert group document on trustworthy AI ([EU HLEG, 2019](#)) and the OECD’s AI principles ([OECD, 2019/2024](#)). Similar analyses and findings were published by [Fjeld et al. \(2020\)](#), [Hagendorff \(2020\)](#), [Ryan & Stahl \(2021\)](#), [Kluge Correa et al. \(2023\)](#).

We used the 11 principles identified by Jobin et al. ([2019](#)) as a starting point for this ontology, using above mentioned analyses for enriching the definition of high-level translational concepts for addressing gaps. These 11 evidence-based principles correspond well to the principles suggested by the European Group on Ethics in Science and New Technologies ([European Commission, 2018](#)) and to the complex of four ethical principles and seven key requirements used by another independent high-level expert group consulted by the European Commission ([EU HLEG, 2019](#)).

In a first step, we carefully mapped the 11 consensus principles against the principles outlined in public documents on trustworthy AI (including those published after the review by Jobin). Not surprisingly, we found that the 11 consensus principles cover the principles outlined in these documents except for topics relating to political values or technological innovation (see below). This analysis is out of scope of this ontology and will be summarised in a forthcoming publication. In a second step, we made minor adaptations based on our premises and observations made in other reviews (section 3), in particular that by Hagendorff ([2020](#)).

We settled finally on **ten principles**:

- **nine ethical principles plus**
- **an overarching aim of ‘trust and trustworthiness’** which we do not consider an ethical principle (no meaningful ethical demand can be formulated), but a *desideratum* that functions as a prerequisite for adoption of and reliance on AI solutions. **Figure 9** summarises how we derived the ten principles. The following sections provide an outline of the approach for defining
- Ethical principles
- Translational concepts branching off from these ethical principles
- Fundamental concepts of related to 12 thematic clusters, required for defining the translational concepts

### 2.1 Premises

In his analysis of AI ethics guidelines and their impact on AI development, Hagendorff also discusses potential *omissions* of topics in a set of ethical guidelines examined ([Hagendorff \(2020\)](#)) and questioning the sufficiency of focusing on technical fixes only: „*In AI ethics, technical artefacts are primarily seen as isolated entities, that can be optimized by experts so as to find technical solutions for technical problems. What is often lacking is a consideration of the wider contexts and the comprehensive relationship networks in which technical systems are embedded*“. Selbst and colleagues have made a similar comment in the context of machine learning fairness ([Selbst et al., 2019](#)).

These excellent reviews served as a basis for outlining consensus ethical principles. This enabled us to define an ontological system of translational concepts for each of the ethical principles.

**Our premises were:**

- **Use of consensus principles identified by previous reviews:** The ethical principles used for an ontological system should be based on the common set of principles identified by the reviews mentioned above and, in particular, by Jobin et al. (2019) who analytically dissected a large body of literature in regard to ethical terms.
- **Granularity:** The ethical principles of the ontology should be ‘granular’, i.e. relating to one ethical demand. ‘Composite principles’, i.e. aggregates of more than one ethical principle and/or technical requirement should be avoided in the interest of clarity.
- **Health relevance:** The ethical principles should capture ethical topics for the health application. Consequently, the principles of ethics in biomedicine were taken into account (→ **bioethics**). Many public ethics guidelines for instance do not include the principle of ‘*beneficence*’, which is critical for health AI applications (e.g. benefit risk assessments, clinical investigations, health technology assessment).
- **Addressing omissions and additional aspects:** Omissions should be analysed in regard to their relevance to health and, if so, addressed either in ethical principles or lower-level translational concepts. Further, additional aspects relating to societal topics (e.g. democratic debate, innovation, public interest) should be considered.

## 2.2 Derivation of nine ethical principles plus trust as a desideratum

**Figure 8** outlines how we arrived at 9 ethical principles based on the 11 principles identified by Jobin et al. (2019):

- **Trust is not an ethical principle, but rather an overarching desideratum and a prerequisite for adoption of AI**<sup>30</sup> (see the discussion under → **TRUST AND TRUSTWORTHINESS** and → **AI principles and AI ethics guidelines**). We therefore denote trust as concept A.0, while the nine ethical principles that are conditions of trust are numbered A.1 to A.9.
- We merged **freedom and autonomy** with **dignity** due to their strong interconnectedness, in particular in the Western philosophy of the enlightenment which is foundational to the concept of human rights and freedoms. For a more detailed explanation see → **DIGNITY, FREEDOM AND AUTONOMY**. Briefly, dignity, i.e. the intrinsic inviolable value of each person can be considered at the root of the **personal freedoms of choice** and **freedom from causation**<sup>31</sup>: freedom (of the will) and personal autonomy are nearly congruent.
- **Fairness and justice:** Under the fairness principle various concepts have been grouped, including the right to challenge, the right for redress and legal liability (Jobin et al. 2019). We moved these latter aspects under the principle of *responsibility*, thus focusing fairness on the ethical requirement (rooted in dignity) to treat people equally, i.e. to avoid discriminatory practices and act in an inclusive manner.

---

<sup>30</sup> See also EU high-level expert group on artificial intelligence: ethics guidelines for trustworthy AI. Online: [https://ec.europa.eu/futurium/en/ai-alliance-consultation\\_1.html](https://ec.europa.eu/futurium/en/ai-alliance-consultation_1.html)

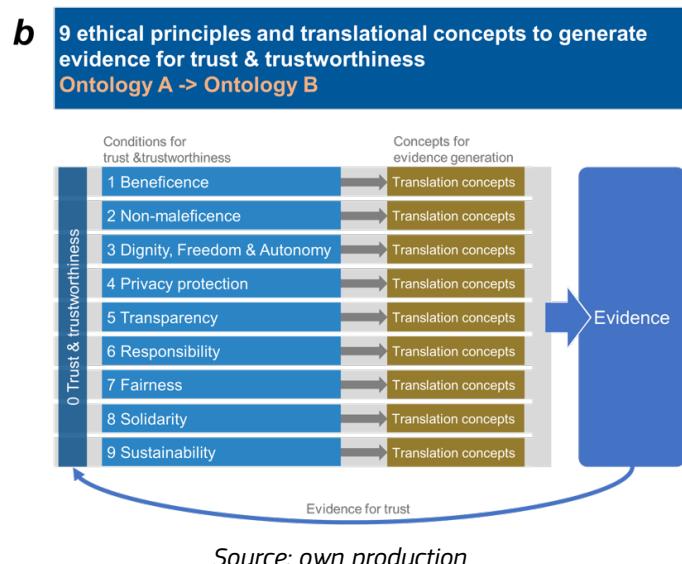
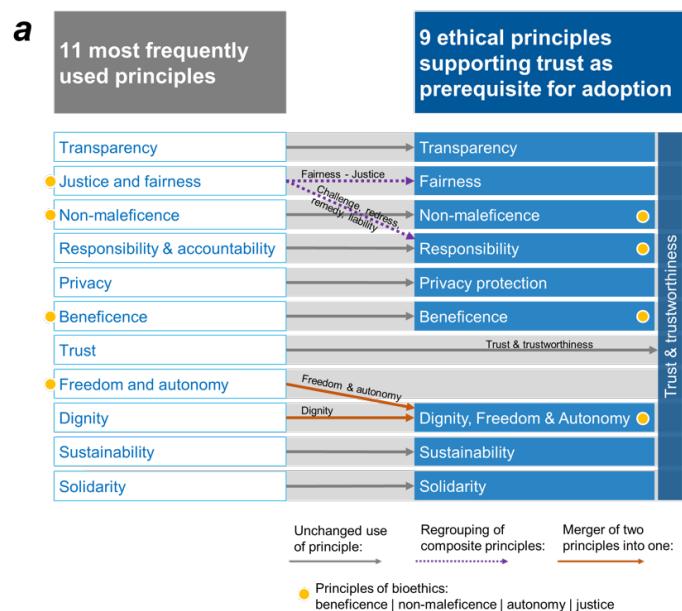
<sup>31</sup> Capes JA (2017) provides a stimulating review on the reconcilability of freedom of will with causation.

Responsibility, a principle that is typically poorly defined (Jobin et al., 2019), includes in our ontology ethical obligations relating to integrity, accountability, human oversight, respect for the body of applicable legislations, but also the entire complex of challenge, contestability, remedy, redress and liability.

In addition, we addressed potential omissions identified by Hagendorff (2020) and included additional aspects outlined in trustworthy AI principles issued by public bodies (**Table 1**). Briefly, these are “respect for democratic values” (OECD, 2019/2024), “public interest” (WHO, 2021) and “safe innovation” (Council of Europe, 2024a, b).

Given the influence of the ethics guideline the European Commission’s independent ‘high-level expert group on artificial intelligence’ for promoting the trustworthy AI concept (EC HLEG, 2019), we summarise the relationship between the group’s seven “key requirements” and ethical principles, translational and fundamental concepts of this ontology (**Table 2**).

**Figure 9. (a)** Derivation of the nine ethical principles of this ontology from the 11 consensus principles identified by Jobin et al., 2019. **(b)** Schematic depiction of the ethical principles and translational concepts for evidence generation to support trust and trustworthiness.



**Table 1.** When developing the ontology we addressed possible omissions of topics in ethical guidelines, based on the analysis by [Hagendorff \(2020\)](#) as well as our own reflections. The table summarizes our reflections. Ethical principles in red font, translational concepts in black font.

| Nr. | Possible omissions*                                     | Specific notions in the health context   | Addressed under ethical principle / translational concept   |
|-----|---|--|---|
| 1   | Prohibitions of use*                                    | Discouraged use scenarios, e.g. coercive practices or 'nudging'  | DIGNITY, FREEDOM AND AUTONOMY <ul style="list-style-type: none"> <li>AI and the development of healthcare - Open, democratic debate</li> </ul>  |
| 2   | Artificial general intelligence                         | N.A.   | N.A.  |
| 3   | Democratic control**                                    | <ul style="list-style-type: none"> <li>Democratic debate about societal consequences (solidarity)</li> <li>Development of healthcare, patient primacy, patient-physician relationship (dignity)</li> </ul> | SOLIDARITY <ul style="list-style-type: none"> <li>Open democratic and solidary society, culture of dialogue</li> </ul> DIGNITY, FREEDOM AND AUTONOMY <ul style="list-style-type: none"> <li>AI and the development of healthcare - Open, democratic debate</li> </ul>   |
| 4   | Social cohesion   | <ul style="list-style-type: none"> <li>Vulnerable groups (e.g. rare diseases)</li> <li>Job impacts in the health sector</li> <li>Economisation of healthcare</li> </ul>                                    | SOLIDARITY <ul style="list-style-type: none"> <li>Taking vulnerable groups into account by design</li> <li>Job impacts, skills, training</li> </ul> DIGNITY, FREEDOM AND AUTONOMY <ul style="list-style-type: none"> <li>AI and the development of healthcare - Open, democratic debate (,Economisation of healthcare ')</li> </ul> |
| 5   | Political abuse of AI                                   | Coercive practices in healthcare and health management   | DIGNITY, FREEDOM AND AUTONOMY <ul style="list-style-type: none"> <li>AI and the development of healthcare - Open, democratic debate (,Coercive practices' / 'nudging ')</li> </ul>  |
| 6   | Lack in diversity                                       | Inclusion, non-discrimination, diversity, plurality: bias  | FAIRNESS<br>NON-MALEFICENCE   |
| 7   | Robot ethics  | Conversational agents / chatbots for health system management or treatment of diseases   | DIGNITY, FREEDOM AND AUTONOMY <ul style="list-style-type: none"> <li>AI and the development of healthcare - Open, democratic debate (see also → conversational agents)</li> </ul>   |
| 8   | Algorithmic decision making better or worse than humans | AI systems for clinical decision making and diagnostics where AI augments human consequential decision making  | BENEFICENCE <ul style="list-style-type: none"> <li>Added value</li> </ul>   |
| 9   | Hidden sociological and ecological cost of AI systems   | <i>Sociological:</i> <ul style="list-style-type: none"> <li>Job losses in health sector, deskilling.</li> </ul>  | SOLIDARITY <ul style="list-style-type: none"> <li>Job impacts, skills &amp; training</li> </ul> DIGNITY, FREEDOM AND AUTONOMY <ul style="list-style-type: none"> <li>Upholding a trustful patient – physician relationship -Deskilling</li> </ul>   |

|  |   | <p><i>Environmental:</i></p> <ul style="list-style-type: none"> <li>• Environmental impact of developing and running AI</li> <li>• Secondary environmental impacts of AI use</li> </ul>   | <b>SUSTAINABILITY</b>   |
|--|---|---|---|
| *) The EU AI Act prohibits specific AI systems. However, more detailed discussions may be needed in application fields such as health regarding use scenarios that should not be encouraged. |   |   |   |
| **) The OECD value-based principles mention democratic values, i.e. principle 3: „ <i>Respect for the rule of law, human rights and democratic values, including fairness and privacy</i> “. |   |   |   |
| Nr.  | <b>Additional aspects</b>                         | <b>Specific notions in the health context</b>   | <b>Addressed under ethical principle / translational concept</b>  |
| 1  | “Respect for democratic values” (OECD, 2019/2024) |   | <b>SOLIDARITY</b> <ul style="list-style-type: none"> <li>• Open democratic and solidary society, culture of dialogue</li> </ul>   |
| 2  | “Public interest” (WHO, 2021)                     | <ul style="list-style-type: none"> <li>• AI in the service of diverse populations</li> </ul>  | <b>SOLIDARITY</b> <ul style="list-style-type: none"> <li>• Open democratic and solidary society, culture of dialogue</li> </ul><br><b>FAIRNESS</b> <ul style="list-style-type: none"> <li>• Avoiding discrimination and discriminatory bias</li> </ul> <p><i>Concerning the tension between interest of groups versus the individuals, see also: DIGNITY, FREEDOM AND AUTONOMY</i></p> <ul style="list-style-type: none"> <li>• Respecting patient primacy</li> </ul> |
| 3  | “Safe innovation” (Council of Europe, 2024).      | <ul style="list-style-type: none"> <li>• Collaborations of agents and communities across life cycle and value chain</li> <li>• Community bridging and relevant tools (e.g. AI evidence pathway)</li> <li>• Education tools to enhance awareness of safety issues</li> </ul> | <b>SOLIDARITY</b> <ul style="list-style-type: none"> <li>• Safe innovation: community bridging, collaboration, education</li> </ul>   |

**Table 2.** Relationship between the seven key requirements of the “ethics guidelines for trustworthy AI” by the EU Commission’s independent high-level expert group (EC HLEG, 2019) and concepts outlined in this ontology. The definitions of key requirements are from Recital 27 of the EU’s AI Act, except for ‘accountability’. Ethical principles (part A) in red font and thematic clusters (part B) in green font. Concepts in black.

| Key requirements (high-level expert group) as defined in EU AI Act (Recital 27)   | Relation to concepts in this ontology.   |
|---|--|
| <p><b>1 Human agency and oversight</b><br/> <i>“means that AI systems are developed and used as a tool that serves people, respects human <b>dignity</b> and personal <b>autonomy</b>, and that is functioning in a way that can be appropriately controlled and <b>overseen by humans</b>.”</i></p>  | <p><b>A.2 NON-MALEFICENCE</b> (in particular)</p> <ul style="list-style-type: none"> <li>• Risks related to dignity, freedom and autonomy</li> <li>• Risks related to responsibility</li> </ul> <p><b>A.3 DIGNITY, FREEDOM AND AUTONOMY</b> (in particular)</p> <ul style="list-style-type: none"> <li>• Respecting patient primacy</li> <li>• AI and the development of healthcare</li> <li>• Upholding a trustful patient-physician relationship</li> <li>• Ensuring the means for free and informed consent</li> <li>• Right to know and right not to know</li> <li>• Right to know if AI system is employed</li> </ul> <p><b>A.6 RESPONSIBILITY</b> (in particular)</p> <ul style="list-style-type: none"> <li>• Ensuring human agency and oversight</li> </ul> <p><b>B.4 Agency, autonomy and automation</b> (in particular)</p> <ul style="list-style-type: none"> <li>• Human agency</li> <li>• Human oversight</li> <li>• Human primacy</li> <li>• Corrigibility</li> <li>• Patient primacy</li> <li>• Automation</li> <li>• Augmentation (in health)</li> </ul> <p><b>B.12 Use of AI systems in health and healthcare</b> (in particular)</p> <ul style="list-style-type: none"> <li>• User research</li> <li>• User competency and training requirements</li> <li>• Instructions for use</li> <li>• Applicability and limitations</li> </ul> |
| <p><b>2 Technical robustness and safety</b><br/> <i>“means that AI systems are developed and used in a way that allows robustness in the case of <b>problems</b> and <b>resilience against attempts</b> to alter the use or performance of the AI system so as to allow unlawful use by third parties, and minimise unintended harm.”</i></p> | <p><b>A.2 NON-MALEFICENCE</b>, in particular:</p> <ul style="list-style-type: none"> <li>• Risks related to insufficient robustness / resilience <ul style="list-style-type: none"> <li>◦ Robustness/Resilience: Use context, use environment variations &amp; drift issues</li> <li>◦ Cybersecurity: risks for value chain assets and knock-on effects on safety</li> </ul> </li> </ul> <p><b>B.2 Ethics, AI ethics, governance, management</b> (in particular)</p> <ul style="list-style-type: none"> <li>• AI risk management</li> <li>• AI safety</li> </ul>   |

| Key requirements (high-level expert group) as defined in EU AI Act (Recital 27)  | Relation to concepts in this ontology.   |
|--|--|
| <p><b>3 Privacy and data governance</b><br/> <i>"means that AI systems are developed and used in accordance with <b>privacy</b> and <b>data protection rules</b>, while <b>processing data</b> that meets high standards in terms of <b>quality and integrity</b>."</i></p>  | <p><b>A.2 NON-MALEFICENCE</b> (in particular)</p> <ul style="list-style-type: none"> <li>Risks related to data privacy of personal information</li> </ul> <p><b>A.4 PRIVACY PROTECTION</b></p> <ul style="list-style-type: none"> <li>Data protection</li> <li>Data security</li> </ul> <p><b>A.3 DIGNITY, FREEDOM AND AUTONOMY</b> (in particular)</p> <ul style="list-style-type: none"> <li>Medical privacy / health privacy</li> </ul> <p><b>B.5 Bias, heuristics, drift &amp; shift</b> (in particular)</p> <ul style="list-style-type: none"> <li>Data drift / shift</li> </ul> <p><b>B.7 Data:</b> (in particular)</p> <ul style="list-style-type: none"> <li>Data provenance of development data</li> <li>Data quality</li> <li>Data quality metrics</li> <li>Synthetic health data</li> <li>Data privacy</li> <li>Personal data</li> <li>CIA principles</li> <li>Data processing / wrangling</li> <li>Data FAIRification</li> </ul> <p><b>B.8 Algorithm, model, algorithm-to model transition</b> (in particular)</p> <ul style="list-style-type: none"> <li>Algorithm-to-model transition</li> </ul> |
| <p><b>4 Transparency</b><br/> <i>"means that AI systems are developed and used in a way that allows appropriate <b>traceability</b> and <b>explainability</b>, while making <b>humans aware</b> that they communicate or interact with an AI system, as well as duly informing deployers of the <b>capabilities and limitations</b> of that AI system and <b>affected persons about their rights</b>."</i></p> | <p><b>A.2 NON-MALEFICENCE</b> (in particular)</p> <ul style="list-style-type: none"> <li>Risks related to transparency</li> </ul> <p><b>A.3 DIGNITY, FREEDOM AND AUTONOMY</b> (in particular)</p> <ul style="list-style-type: none"> <li>Right to know if AI system employed</li> </ul> <p><b>A.5 TRANSPARENCY</b></p> <ul style="list-style-type: none"> <li>Transparency of organisation and actors providing / deploying AI systems</li> <li>Transparency of human-AI interaction</li> <li>Traceability</li> <li>Failure transparency</li> <li>Transparency of AI systems: evidence needs (all lower-level translational concepts)</li> </ul>   |

| Key requirements (high-level expert group) as defined in EU AI Act (Recital 27)  | Relation to concepts in this ontology.  |
|--|---|
| <p><b>5 Diversity, non-discrimination and fairness</b><br/> <i>"means that AI systems are developed and used in a way that includes <b>diverse actors</b> and promotes <b>equal access, gender equality and cultural diversity</b>, while avoiding <b>discriminatory impacts</b> and <b>unfair biases</b> that are prohibited by Union or national law."</i></p> | <p><b>A.2 NON-MALEFICENCE</b> (in particular)</p> <ul style="list-style-type: none"> <li>• Risks related to bias</li> </ul> <p><b>A.7 FAIRNESS</b></p> <ul style="list-style-type: none"> <li>• Avoiding discrimination and discriminatory bias – fairness and underrepresentation of groups in AI development data sets <ul style="list-style-type: none"> <li>○ Looking out for and avoiding discriminatory bias</li> <li>○ Monitoring and mitigation of possible discrimination throughout the evidence pathway</li> </ul> </li> <li>• Health equality and health equity</li> <li>• Unavoidable trade-offs</li> <li>• Universal versus targeted design</li> </ul> <p><b>A.8 SOLIDARITY</b> (in particular)</p> <ul style="list-style-type: none"> <li>• Taking vulnerable groups into account "by design"</li> </ul> <p><b>B.5 Bias, heuristics, drift &amp; shift</b> (in particular)</p> <ul style="list-style-type: none"> <li>• Bias</li> <li>• Heuristics</li> <li>• Intrinsic incompatibilities or 'trade-offs'</li> </ul> |

| Key requirements (high-level expert group) as defined in EU AI Act (Recital 27)   | Relation to concepts in this ontology.   |
|---|--|
| <p><b>6 Social and environmental well-being</b><br/> <i>"means that AI systems are developed and used in a sustainable and environmentally friendly manner as well as in a way to benefit all human beings, while monitoring and assessing the long term impacts on the individual, society and democracy."</i></p> | <p><b>A.3 DIGNITY, FREEDOM AND AUTONOMY</b> (in particular)</p> <ul style="list-style-type: none"> <li>• AI and the development of healthcare <ul style="list-style-type: none"> <li>◦ Open, democratic and evidence-based debate</li> </ul> </li> </ul> <p><b>A.8 SOLIDARITY</b></p> <ul style="list-style-type: none"> <li>• Open democratic and solidary society, culture of dialogue</li> <li>• Taking vulnerable groups into account "by design"</li> <li>• Accessibility of data and infrastructure for AI development</li> <li>• Avoiding colonial structures; reducing global north-south divide</li> <li>• Job impacts, skills, training</li> <li>• Safe innovation: community bridging, collaboration, education</li> </ul> <p><b>A.9 SUSTAINABILITY</b></p> <ul style="list-style-type: none"> <li>• Environmental sustainability of AI systems throughout the life cycle</li> <li>• Sustainability impact of AI systems on health systems</li> </ul> <p><b>B.2 Ethics, AI ethics, governance, management</b> (in particular)</p> <ul style="list-style-type: none"> <li>• AI impact assessment (AI-IA)</li> <li>• Fundamental rights and algorithm impact assessments</li> </ul> |

| Key requirements (high-level expert group) as defined in EU AI Act (Recital 27)  | Relation to concepts in this ontology.  |
|--|---|
| <p><b>7 Accountability*</b><br/> <i>“... is closely linked to the principle of fairness. It necessitates that mechanisms be put in place to ensure responsibility and accountability for AI systems and their outcomes, both before and after their development, deployment and use.” (EU HLEG, 2019, page 19)</i></p> | <p><b>A.2 NON-MALEFICENCE</b>, in particular:</p> <ul style="list-style-type: none"> <li>Risks related to responsibility</li> </ul> <p><b>A.6 RESPONSIBILITY</b>, in particular:</p> <ul style="list-style-type: none"> <li>Legal and regulatory compliance</li> <li>Acting with integrity <ul style="list-style-type: none"> <li>Ethics code and governance / management framework</li> <li>Quality culture and risk management</li> <li>Correcting problems and failures, including necessary communication</li> <li>Peer review and community discourse</li> </ul> </li> <li>Accountability <ul style="list-style-type: none"> <li>Accountability structures, attribution of (distributed) responsibilities</li> <li>Auditability and auditing</li> </ul> </li> <li>Responsiveness, contestability, redress, liability</li> </ul> <p><b>B.2 Ethics, AI ethics, governance, management</b>, in particular:</p> <ul style="list-style-type: none"> <li>AI governance</li> <li>AI management</li> <li>AI risk management</li> </ul> <p><b>B.3 AI actors and communities</b> (all concepts)</p> <p><b>B.6 AI Evidence pathway, AI life cycle and AI value chain</b>, in particular:</p> <ul style="list-style-type: none"> <li>Post-deployment monitoring</li> <li>Decommissioning / retirement</li> </ul> <p><b>B.9 Relevance</b> (all concepts)</p> <p><b>B.10 Verification and validation</b> (all concepts)</p> <p><b>B.11 Clinical evidence, evaluation, validation</b>, in particular:</p> <ul style="list-style-type: none"> <li>Clinical evaluation</li> <li>Clinical validation</li> <li>Clinical investigation</li> <li>Post-market surveillance, market surveillance, corrective action</li> <li>Usability</li> </ul> |

\*) Recital 27 does not provide a definition of ‘accountability’, the seventh key requirement proposed by the EU high-level expert group in its 2019 guidance document.

## 2.3 Fundamental concepts associated with ethical principles and translational concepts

The fundamental concepts (i.e. entities or terms) elaborated in part B of this ontology were defined through a top-down approach: by first defining ethical principles and unfolding these

into translational concepts, we identified necessary ‘fundamental concepts’ of relevant technical, scientific, ethical, philosophical, clinical and (health)economic terms. All terms are underpinned by relevant references. In a final step, the fundamental concepts were arranged in 12 clusters so as to enable users of this ontology to explore thematically related topics in a more convenient manner.

### 3 Ontology A: Ethical principles and translational concepts

#### A.0 Trust and trustworthiness

*"Through the trust which a person either shows or asks of another person, he or she surrenders something of his or her life to that person. Therefore, our existence demands of us that we protect the life of the person who has placed his or her trust in us."*

Løgstrup, KE "The ethical demand" ([Løgstrup, 1956](#))

##### Concept description

**Trust is an essential part of human relations as well as reliance on technologies. This holds particularly for AI: its unique capacity to produce outcomes that, until now, required uniquely human skills, requires trust in the actors developing AI and in AI technology itself. Trust is particularly important in health and healthcare, due to the fundamental importance of health for living our lives and the vulnerability of being a patient.**

Not surprisingly, trust is mentioned in many AI guidelines. While such documents (see **Box 2**) may stipulate conditions for trust (see section 2.1 below), the concepts of 'trust' and 'trustworthiness' are typically not further expounded. We hence start with a brief exploration of the concepts 'trust and trustworthiness'.

##### 1 Trust and technologies

Trust is not an ethical principle: no ethical demand for trust can be formulated. Instead, trust can be defined as a manifestation of faith, belief or confidence in *something* or *someone* and may be associated with positive notions such as, reliability, value, truth or truthfulness, good intentions, ability, usefulness, skill, capacity etc. ([Mayer et al., 1995](#); [Hancock et al., 2023](#)). In his seminal work "The ethical demand" [Løgstrup \(1956\)](#) introduced the notion that basic trust is a fundamental requirement for and currency of human interactions that cannot itself be deduced from a rule or law, including deontic ethical principles or ethical demands (see → **ethical principles**). However, the act of trusting implies ethical demands to protect that very (a priori) trust (see citation above). This can be done by stipulating 'conditions for trust' such as identifying ethical principles or demands.

Trust and trustworthiness have long been an issue when new technologies became available, including trust in medical technologies and biomedicine (see for instance: [Enid et al., 2009](#)). The debate around → **bioethics** in the 1970ies led to the formulation of four basic '*bioethical*' principles (see section 2, below)<sup>32</sup>. Inversely, *unjustified* trust in automation and algorithms has also been subject to intense debate. This is exemplified in the AI field by the early chatbot experiment ELIZA, created by one of the AI pioneers Joseph Weizenbaum in the 1960ies. ELIZA was, by his own account and to his dismay, adopted by some psychologists for treating patients (see: [Campolo et al., 2016](#) and [Prujit, 2006](#)).

Trust as a key aspiration or "principle" is mentioned in various → **AI ethics** documents, including in context of 'data trust' and 'public trust' ([Jobin et al., 2019](#); [Hagendorff, 2020](#); [Fjeld et al., 2021](#); [Ryan & Stahl, 2022](#)). Trust received particular attention with the publication of widely disseminated documents, notably the EU Commission's ethics guidelines on "trustworthy AI", prepared by a group of experts ([EC HLEG, 2019](#)) or Price, Waterhouse & Cooper's (PWC) "A practical guide to responsible AI" ([PWC, 2019](#)). The EU high-level expert group used three of the four bioethical principles, complemented by another

<sup>32</sup> Some researchers have suggested to restrict the concept of trust for interpersonal matters (including trust in organisations), suggesting to rather use the concept of "reliance" ([Deley & Dubois, 2020](#)) in the context of technologies.

principle termed → **explicability**. The group introduced seven ‘key requirements’ for realising these ethical principles and to achieve ‘trustworthy AI’ ([EU HLEG, 2019](#)). One of the bioethical principles, “*beneficence*”, was not included in the expert group’s document<sup>33</sup>.

## **2 Trustworthy AI: ethical principles and legislations**

The ethics guidelines on trustworthy AI by the EU Commission’s expert panel ([EU HLEG, 2019](#)) have influenced initiatives and regulatory debate on a global level, for example the US Executive Order on Safe, Secure, and Trustworthy AI ([White House, 2023, rescinded in 2025](#)), the Bletchley Declaration on AI ([UK government, 2023](#)) and the Paris AI Action Summit ([French Government, 2025](#)).

The HLEG guidelines, among a series of academic papers (e.g. [Floridi et al., 2018](#); [Floridi & Cowls, 2019](#)) contributed to a deepened debate on what ethical concerns need to be considered to ensure that the end product will merit trustworthiness from an ethical, societal and also technical perspective. While the debate (in particular in the EU) first centred around ethical principles and key requirements formulated as conditions for trustworthy AI, it is clear that AI systems pose specific challenges that require also clear legislative and regulatory approaches.

AI regulations will be key to shaping robust AI governance frameworks for trustworthy AI on a global level, in particular from a product safety perspective and one of conformity with essential requirements. The EU’s AI Act ([EU, 2024a](#)) covers AI applications also in healthcare. Annex III refers to essential public services and emergency services including healthcare, while Annex I points to relevant other EU harmonisation legislation, such as the MDR and IVDR. Using AI in health in a responsible manner will certainly require a broad and people-centric approach: “...the primary forces that are needed to unlock the value from artificial intelligence are people-based and not technical.” ([OECD, 2024d: ‘Collective action for responsible AI in health’](#)).

**Taken together, the concept of trust in AI in the health domain involves two broad avenues: soft law and legislations**

- 1) **ethical (‘soft law’) framing conditions for trust** that are translatable into **actionable topics tailored to a field of application**. In health this includes
  - a. **aspects of adoption and technical integration into → use environments** (e.g. specific hospital or clinical settings) and → **use contexts**.
  - b. **implementation into clinical workflows**, including skills, continuous training, correct use of AI systems, their monitoring and considerations of shifts or drifts that may alter or degrade their performance)
  - c. discussions concerning the **future of medical care and the role of AI in it**
  - d. the impact of AI on the **patient-physician relationship** (see → **DIGNITY, FREEDOM AND AUTONOMY**), skills of healthcare professionals and
  - e. trust issues related to the insufficient handling of potential **conflicts of interests** originating e.g. from situations where AI developers may at the same time be users in a given environment (e.g. clinical teams developing AI systems for radiological image analysis with a view to their subsequent routine use either within a local framework or as a marketed product).
- 2) **a rule-based (e.g. legislative) framework providing *inter alia* clarity on rules and requirements for AI products based on the risk these pose as well as legal obligations**

<sup>33</sup> Notably, in 2018 the AI4people group ([Floridi et al., 2018; for more details see Floridi & Cowls, 2019](#)) had suggested to use all four ethical principles of bioethics for AI, including *beneficence*. These were complemented by a new principle called ‘**explicability**’, denoting a composite of the epistemological issue of *intelligibility* (how does it work?) with the ethical demand of *accountability*.

**of various actors.** Some of the rules will reflect ethical dimensions. In the EU there is a comprehensive set of legislations relevant in this context, e.g. AI Act, GDPR, Data Act, Cyber Resilience Act etc. A short **overview of these EU legislations in the context of health can be found in the study on “Health Data, Digital Health and Artificial Intelligence in Healthcare”** ([Lupiáñez-Villanueva et al., 2022](#)), commissioned by the European Commission (see in particular table 24 in section 3.2.1).

These two aspects are not separable. The long-standing debate on AI (ethical) principles has in the meantime led to legislations (EU’s AI Act, South Korea’s AI Act) or official decrees (e.g. [US government, White House 2023](#); rescinded in [2025](#)) that are based on AI ethical principles.

### **3 Four dimensions of trust and trustworthiness of AI in health**

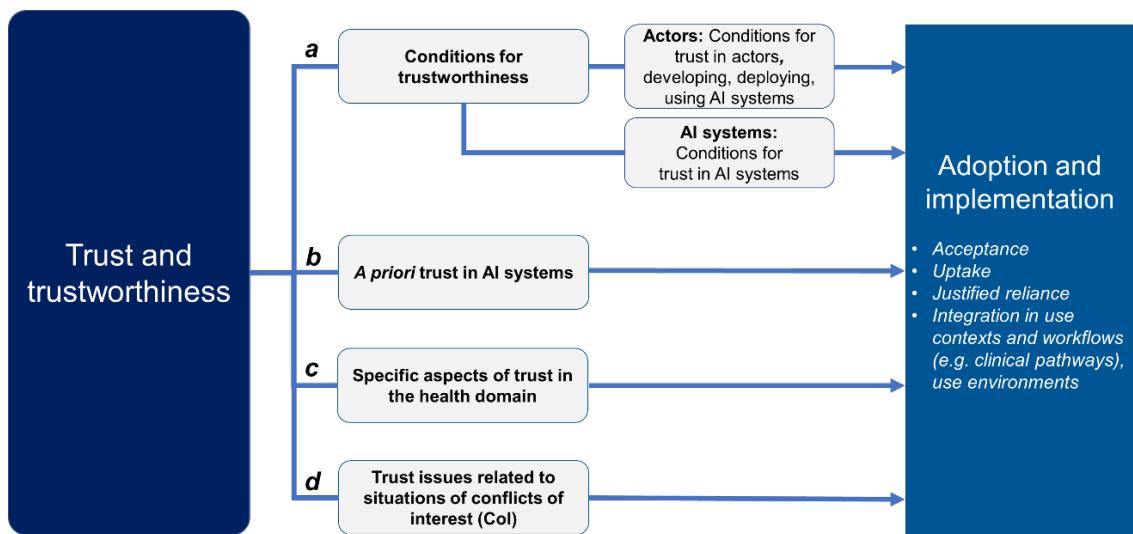
We propose four basic dimensions for trust and trustworthiness of AI in health (**Figure 9**). These are discussed as translational concepts in the sections below.

- a) Conditions for trustworthiness concern both the actors responsible for developing, deploying, using, monitoring, decommissioning AI and the AI systems themselves (i.e. the manner in which a specific AI solution has been designed and developed). Conditions for trustworthiness are based on ethical principles that need to relate to concrete actionable concepts: the purpose of this ontology is to lay out this relationship.
- b) A priori trust in AI systems which can have harmful effects, such as automation bias and complacency (e.g. [Challen et al., 2019; Arnold, 2021](#)). See → avoiding automation bias, → avoiding automation complacency.
- c) Specific aspects of trust in the health domain, e.g. pertaining to potential destructive consequences on the patient-physician relationship (for a comprehensive analysis, see [Mittelstadt, 2021](#)), deskilling of healthcare professionals, “mechanisation” of healthcare, biases, health equality and equity. These are explored under specific ethical principles in this ontology.
- d) Trust related to situations of conflict of interests, e.g. where healthcare professionals are AI developers and, at the same time, have role in deciding on procurement and/or use of AI systems.

### **4 Looking forward**

Building trust in AI technologies is a complex endeavour. A key issue is that aspects central to trust are difficult to define, let alone measure. This implies a certain vagueness associated with trust. As we have proposed, a way forward might be for researchers and stakeholders to explore the tension between evidence needs for trust and realisability of such evidence. Such debate should also address how relevant evidence needs may change or evolve with the rapidly developing AI technology (see also the emerging research field of “predictable AI” ([Future of Life Institute, 2024b; Zhou et al., 2023](#))).

**Figure 10.** Four dimensions of trust as a prerequisite for adoption of AI technology in health. **(a)** Conditions for trustworthiness (=9 ethical principles), concerning both actors involved in building and deploying AI systems and the technology itself. **(b)** A priori trust in AI systems without sufficient oversight (e.g. → automation bias). A priori trust is comparably little debated. **(c)** Specific notions of trust in the context of health applications. This touches on a variety of issues, e.g. privacy of health data, patient-doctor relationship, and informed consent. **(d)** situations of conflicts of interests, e.g. in case roles of developer and decision-maker concerning the acquisition and implementation of AI systems in a given healthcare setting coincide.



Source: own production

#### Explanatory note

Given above considerations, trustworthiness in AI (as for other technologies) will depend on both,

- **legislative frameworks and associated regulatory guidance**
  - to ensure that products are undergoing risk-level dependent scrutiny and fulfil relevant performance requirements (which have been criticised as 'proxies for trustworthiness'; [Laux et al., 2024](#)) to ensure, essentially that there are safe to use and
  - to ensure that products do not violate relevant legislations, treaties and conventions relating to human rights)
- **"soft law" in the form of ethical guidelines** (typically → [principism](#)-based approaches) and **technical, scientific and clinical guidance** to support
  - Communication, education and training of healthcare professionals in view of pitfalls and risks of using AI in health and promote discussion of how to implement AI in health and medicine, e.g. in clinical workflows.
  - common understanding (including between regulatory authorities)
  - global dialogue on challenges concerning health equality and health equity, resulting from an uneven introduction of AI around the globe
  - the further development and refinement of using AI in a responsible manner, in particular of ethical and scientific topics that are out of scope of legislative approaches which are fundamentally aimed at product safety (e.g. EU Regulations on medical devices and in vitro diagnostic medical devices).
  - continuous scientific, ethical and societal discourse in response to new challenges resulting from technological, societal and other changes that relate to fairness; solidarity; dignity, freedom and autonomy;

Finally, trust and trustworthiness are **preconditions for acceptance** (both societal and within specific communities, e.g. by clinicians and patients) and widespread **adoption** of AI. We alert to two publications, relevant in this context:

- Stevens and Stetson (2023) have proposed a **model to measure and explain trust and adoption of AI technology in clinical settings** that may be informative for various actors along the → life cycle of AI in health; → AI evidence pathway for health.
- Tamò-Larrieux and colleagues (2024) have approached the issue of **trust in AI from the perspective of trust in automation** and suggest **16 propositions concerning trustworthiness**. Many of these take **context-dependent aspects into account** that are often insufficiently considered (see also Selbst et al., 2019). Some of these overlap with relevant legislations (e.g. EU' AI Act); EU, 2024a), others are not necessarily reflected in legislative approaches. Such aspects are however of high relevance when considering the integration of AI into **clinical workflows**, which is not a matter of product safety but of decisions about best practices within given → use environments and for given → use contexts which are highly adaptable.

Both proposals are an excellent basis for further discussion concerning **trust in and adoption of AI in health and medicine**.

#### Term relationship

Related terms:

- **Principism**

#### Ethical principles used in this ontology to realise trust in AI:

- BENEFICENCE
- NON-MALEFICENCE
- DIGNITY, FREEDOM AND AUTONOMY
- PRIVACY PROTECTION
- TRANSPARENCY
- RESPONSIBILITY
- FAIRNESS
- SOLIDARITY
- SUSTAINABILITY

**Box 2.** Trust and trustworthiness are part of most documents on AI issued by public organisations

Various international and national **public organisations** have issued approaches for AI, all of these but one ([EGE, 2018](#)) refer to trust and trustworthiness. All are making use of → **principism** introduced with → **bioethics**, i.e. they formulate ethical or value-based “principles” or “characteristics” ([NIST, 2023](#)) to group and communicate relevant matters (→ AI principles and AI ethics guidelines):

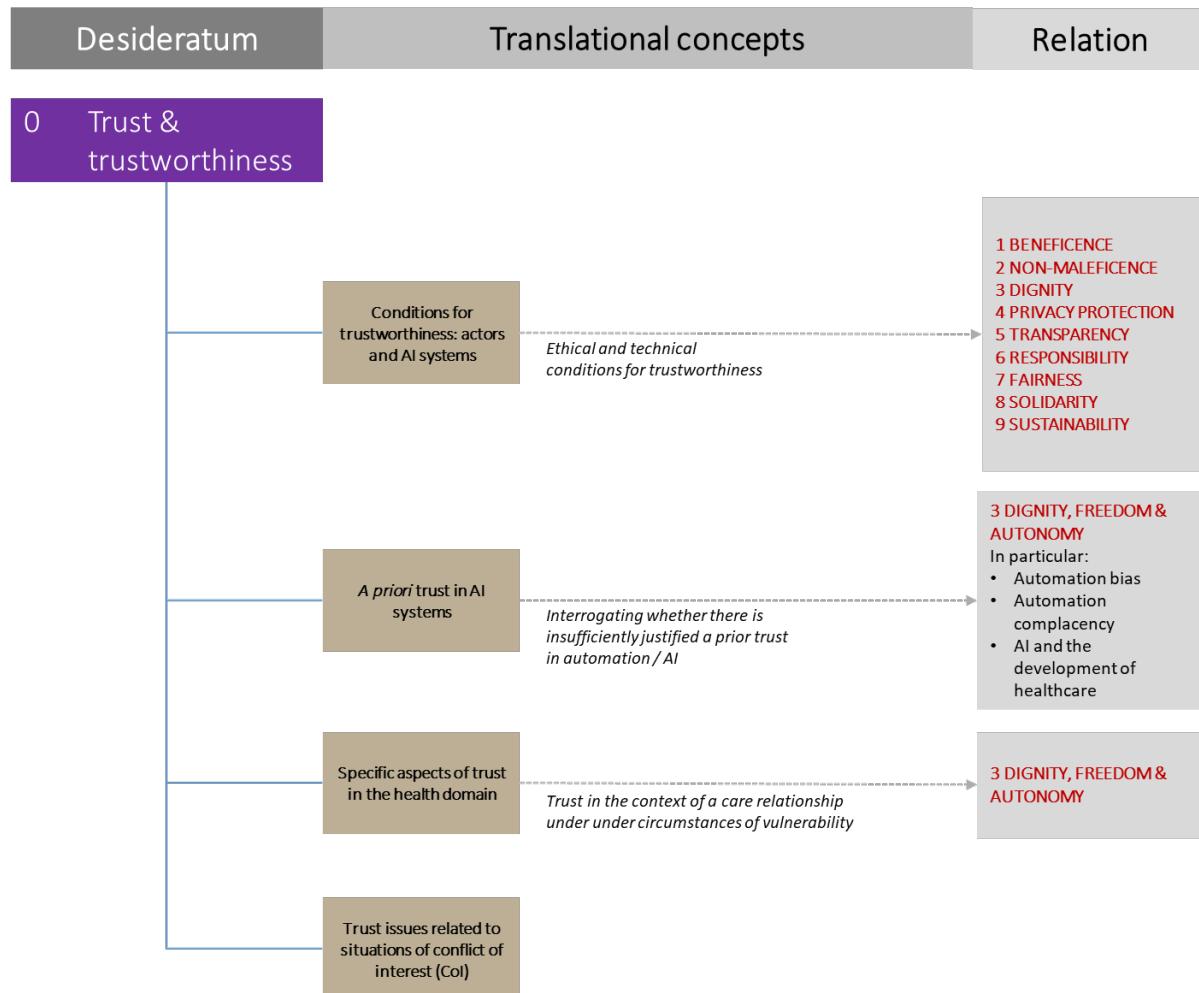
- The **EU Commission expert group on ethics in science and new technologies** ([European Commission, 2018](#)) issued a statement AI, automation and robotics, using nine ethical principles to structure relevant ethical issues. These principles are rooted in the values laid down in EU treaties and in the EU Charter of fundamental rights. These treaties are foundational for the EU’s “AI Act” which came into force 1 August 2024 ([EU, 2024a](#)). The document calls for a ...”common, *internationally recognised ethical ... framework for the design, production, use and governance of artificial intelligence, robotics, and ‘autonomous’ systems.*” Trust is not mentioned in this guidance.
- The **EU Commission’s independent high-level expert group** proposed a framework for trustworthy artificial intelligence, using **four ethical principles** as a foundation for **seven “key requirements”** which are a blend of ethical principles ([Jobin et al., 2019](#)) and technical concepts.
- In 2021, an expert panel working on request of the European Commission’s DG Research and Innovation issued a document on “Ethics by design and ethics of use approaches for artificial intelligence, listing **six ethical principles**, which were “*informed by the work of the Independent High-Level Expert Group on AI (AI-HLEG) set up by the European Commission. They are also based on value frameworks proposed by the European Group on Ethics in Science and New Technologies, Statement on Artificial Intelligence, Robotics and ‘Autonomous’ Systems, 2018, the Institute of Electrical and Electronics Engineers (IEEE), the Organisation for Economic Co-operation and Development (OECD) and UNESCO.*” and “...also based on value frameworks proposed by the European group on ethics in science and new technologies.” The document thus implicitly refers to trust and trustworthiness.
- The **Organisation for Economic Development and Cooperation** ([OECD 2019; 2024](#)) refers to its principles as “value-based principles” and mentions trustworthiness as a goal. Moreover, the OECD has published a **framework to compare implementation tools for trustworthy AI** from private and public actors, including **procedural, technical and educational tools** ([OECD, 2021b](#)).
- The **WHO’s comprehensive guidance document on ethics and governance for AI** ([WHO, 2021a](#)) on the background of the WHO’s Astana declaration outlining five principles for the use of digital technology ([WHO, 2018](#)) and the WHO global strategy on digital health ([WHO, 2021c](#)). The document refers on many occasions to ‘trust’ and ‘trustworthiness’. In 2024 the WHO provided, in 2024, guidance on ethics and governance specifically of large multi-modal models in health ([WHO, 2024a](#)).
- The **US Department of Health & Human Services** ([United States, 2021](#)) has published a Trustworthy AI (TAI) playbook based on six principles for trustworthy AI proposed by Deloitte’s ‘Ethics of AI framework’ (‘Trustworthy AI™’; [Deloitte, 2022](#)).
- The **US National Institute for standardisation and technology** (NIST, 2023) in its document entitled “Artificial Intelligence Risk Management Framework” ([NIST, 2023](#)) refers to trustworthiness and maps how other key topics (e.g. → AI safety, explainability, → privacy protection) map on trustworthiness.
- The **Council of Europe ‘Convention on artificial intelligence and human rights, democracy and the rule of law’** ([Council of Europe, 2024a](#)) refers to a set of seven

principles that “trustworthy artificial intelligence systems will embody” ([Council of Europe, 2024b](#)).

It is noteworthy that some publications on AI principles and related guidance have opted for a mix of ethical principles and more practice-oriented “requirements”, merging ethical principles with technical requirements as new “principles”. For instance, the EU high-level expert group’s ethics guidelines present three elements under the central tenet of “trustworthy AI”: 1) AI must be lawful, 2) AI must be ethical and 3) AI must be robust. The document outlines four ethical principles and seven key requirements. Some of the “requirements” stipulated can be seen as ethical principles themselves (e.g. diversity, non-discrimination, fairness, transparency) or are rooted in ethical principles (e.g. human agency and oversight, societal and environmental well-being, accountability), while others are rather technical requirements (e.g. technical robustness). Similarly, the FUTRE-AI guideline for trustworthy AI uses six “guiding principles”, which are a blend of ethical principles (e.g. ‘fairness’, ‘universality’ – as a novel umbrella term for equity and equality) and technical requirements (e.g. robustness, explainability, traceability, usability) ([FUTURE-AI, 2025](#)).

## Trust and trustworthiness: translational concepts

Ontological organisation of the desideratum of “trust and trustworthiness”, its translational concepts and relation to other ethical principles



## Conditions for trust and trustworthiness: actors and AI systems

### Parent term: Trust and trustworthiness

In this ontology, we consider **evidence on the nine ethical principles and associated translational concepts** as conditions for trust:

- BENEFICENCE (identifying and monitoring benefits, e.g. for the derivation of a benefit-risk ratio which supports assessing acceptability of residual risks).
- NON-MALEFICENCE (e.g. risk identification to ensure → AI safety and, more specifically, patient safety)
- DIGNITY, FREEDOM AND AUTONOMY (e.g. patient-physician relationship, deskilling)
- PRIVACY PROTECTION (e.g. privacy preserving techniques to enhance trust)
- TRANSPARENCY
- RESPONSIBILITY
- FAIRNESS (avoidance of bias)
- SOLIDARITY (e.g. workplace aspects, not leaving vulnerable groups behind)
- SUSTAINABILITY

Without adequate evidence on topics relating to the ethical principles, there will be insufficient trust. This will hamper adoption and uptake of AI in health.

### Explanatory note

Being essentially a belief or desideratum, trust carries significant risks for the trustor (who lends trust) and obligations for the trustee (who is being trusted). For something to be deemed “trustworthy” (i.e. meriting our trust) the trustor requires **evidence that lends support to the appropriateness, plausibility and justifiability of placing trust in something or someone**, e.g. an organisation, a process, a system, a machine.

In this sense, trust can be described as the outcome of a **cognitive process** (Mayer et al., 1995; see also Tamò-Larrieux et al., 2024; Hancock et al., 2023) that is coupled to **conditions for trust or trustworthiness**. These can be phrased in **technological terms** (e.g. “robustness”) or in **ethical terms** (e.g. “non-maleficence”, “fairness” etc.).

A simple model of trust that has been suggested by Mayer et al. (1995; see also Hancock et al. 2023) which is useful in the present context. The model is composed of four elements: a) perceived factors of ‘*trustworthiness*’ (e.g. benevolence, ability, integrity) of an actor, b) the trustors ‘*propensity*’ to lend trust to that actor, the perceived *risk*, c) the establishment of a *risk-taking relationship* (for technologies such as AI this may also concern considerations of ‘acceptable risks’) and d) the observation of *outcomes* that feed back into notions of trustworthiness.

In the context of health, trust relates to actions, decisions, recommendations of a specific actor. This could be a physician, a method, machinery or test (e.g. for diagnosis) or a → **AI system** – or a blend of these.

## A priori trust in AI systems

### **Parent term:** Trust and trustworthiness

While there has been a focus on how to achieve trust in AI technology through a ‘programme’ of “trustworthy AI” and associated high-level ethical principles and technical concepts, there has been comparatively little debate about the risks of **“*a priori* trust” in AI, i.e. insufficiently justified trust in automation and AI.**

Examples for *a priori* trust are automation bias and complacency (see → avoiding automation bias and → avoiding automation complacency). *A priori* trust has much less to do with product safety than with the manner in which AI is used in practice and implemented in workflows.

*A priori* trust constitutes risks for the patient-physician relationship ([Mittelstadt, 2021](#)). It also poses a risk for the safe use of AI in health ([Challen et al., 2019; Arnold, 2021](#)).

Thus, → AI safety is not only a responsibility of the actors (e.g. companies and, where applicable, third parties assessing the technology prior to deployment, but depends also on downstream actors and → users of AI in the health domain, deciding on how AI is deployed and integrated into (clinical) workflows.

## Specific aspects of trust in the health domain

### **Parent term:** Trust and trustworthiness

Trust may also have a dimension that is specific to the health domain.

In health, trust concerns a **complex nexus of specific vulnerabilities and uncertainties** (e.g. of patients in healthcare or of students in education), of **human relationships** (e.g. patient-physician relationship, pupil-teacher relationship) and the **use, implementation of and reliance on technology, in particular where it augments human agency** (e.g. clinical decision-making systems based on AI, image analysis for diagnostics, robotic surgery) and, in particular, where there are risks that augmentation blends over time into substitution (e.g. due to → automation complacency or → automation bias).

In healthcare **patient trust rests to a great deal on interpersonal interaction** ([Brown, 2009](#)) and technology may erode or undermine such trust in various ways ([Mittelstadt, 2021](#)). Patients are generally *a priori* less trustful towards technological systems ([Pew, 2023](#)), which also underlines the importance of providing patients with understandable information of how AI works and why a specific outcome was produced by an AI system, i.e. for → ensuring the means for free and informed consent. This relates to the necessity to have either models with intrinsic interpretability (e.g. based on a decision-tree or causal learning) or attain interpretability through post-hoc explainability techniques (e.g. in case of deep-learning / black-box models) (see → interpretability and explainability).

The specific aspects of trust in the health domain are mainly covered under → DIGNITY, FREEDOM AND AUTONOMY.

## Trust issues related to situations of conflict of interest (CoI)

### **Parent term:** Trust and trustworthiness

In a recent systematic review concerning barriers of AI implementation in healthcare, Ahmed et al. ([2023](#)) have found that a significant number of studies cited **conflicts of interest** (COI) as a main hurdle for AI uptake – the others being trust in general, → PRIVACY PROTECTION and patient consent (→ ensuring the means for free and informed consent).

It is common (and not *a priori* undesirable), that clinicians (e.g. radiologists) are involved in developing a new AI system for improved diagnosis (see for instance: [Chassaigne et al., 2024; 2025](#)) and even may re-

ceive remuneration/royalties for advisory or other services or be engaged in commercial enterprises. However, there is clearly a potential conflict of interest if such developers, in their functions at a given healthcare settings, are also involved in decisions concerning the implementation of this AI system in that very setting (e.g. a given hospital) or even a health system. Brady et al. (2020) recommend that such COIs must be handled analogously to established practices in relation to development of medicinal products or medical devices.

In addition, COIs may not only be of a commercial nature. Clearly, a clinician who developed an AI system may be biased in regard to its perceived benefits. We strongly recommend a system of → peer review and **community discourse** to ensure the use of objective evidence for any possible decision making. Peer review, however, will not sufficiently address real-world socio-economic benefits of AI system as required for reimbursement decisions (→ life cycle of AI in health; → AI actors and communities – health technology assessment (HTA) experts).

Medical associations could play a role of ensuring, among their members, awareness of issues related to professional integrity, potential conflicts of interests which may affect also patient safety. This could be achieved through guidelines, regular alerts to maintain sensitivity or the formulation of ethics codes that address also conflicts of interests stemming from the potential overlap of roles of developer and user.

## A.1 Beneficence

### Concept description

Beneficence is one of the ethical principles used in → bioethics and clinical ethics. It captures the moral “obligation to act in a manner which confers benefit to others” (Jankowski, 2024). Beneficence is an intrinsic element of clinical medicine concerning all aspects related to *patient benefits* (Beauchamp & Childress, 1979).

Beneficence is of particular importance for AI systems with a medical purpose. A key concept in this context is → clinical benefit. Considerations related to beneficence and benefits are key for → AI risk management of AI systems used in healthcare: acceptability of remaining risks of devices is judged on the basis of weighing risks against benefits (IMDRF, 2018; EU medical devices Regulation, Annex I, Chapter I; EU, 2017). This is usually done through gathering evidence on the benefit-risk ratio during → clinical evaluation.

We use the term beneficence with a broader scope, referring to **benefits for patients as well as individuals and organisations working for the public interest**. Beneficence in the present context includes:

- a) **Benefits to people:** primarily patients and healthcare (→ users of AI in the health domain: healthcare professionals, care staff, care givers). We suggest to consider also benefits for researchers, public health officials and health system manager for contexts outside healthcare, as well as benefits for (patient) groups, for communities and society as a whole.
- b) **Benefits to organisations:** organisations involved in healthcare (e.g. healthcare settings, healthcare systems), organisations conducting health research, public health surveillance for the public interest.

Consequently, benefits for private organisations that do not primarily work for the public good are out of scope.

### Explanatory note

Beneficence is of key importance in biomedicine. Nevertheless, beneficence is typically not included as an ethical principle in public documents outlining AI principles or it is blended with other principles (which we refer to as “composite principles”) and referred to under terms such as “well-being” (e.g. “*inclusive growth, sustainable development and well-being*”, OECD, 2019/2024) or “*promote human well-being*” (WHO, 2021a).

In addition to the physician’s obligation to provide benefit (beneficence) to patients, bioethics and clinical ethics contain the equally important obligations that physicians need to avoid or minimize harm to patients (→ NON-MALEFICENCE), and need to respect the values and preferences of their patient (→ DIGNITY, FREEDOM AND AUTONOMY, → FAIRNESS).

Beneficence should not be confused with benefit or → clinical benefit. Beneficence is the ethical obligation to create benefits, including → clinical benefits in case of AI systems for medical purposes. Benefits in contrast are observable and measurable positive impacts or outcomes.

For additional references, see the following terms:

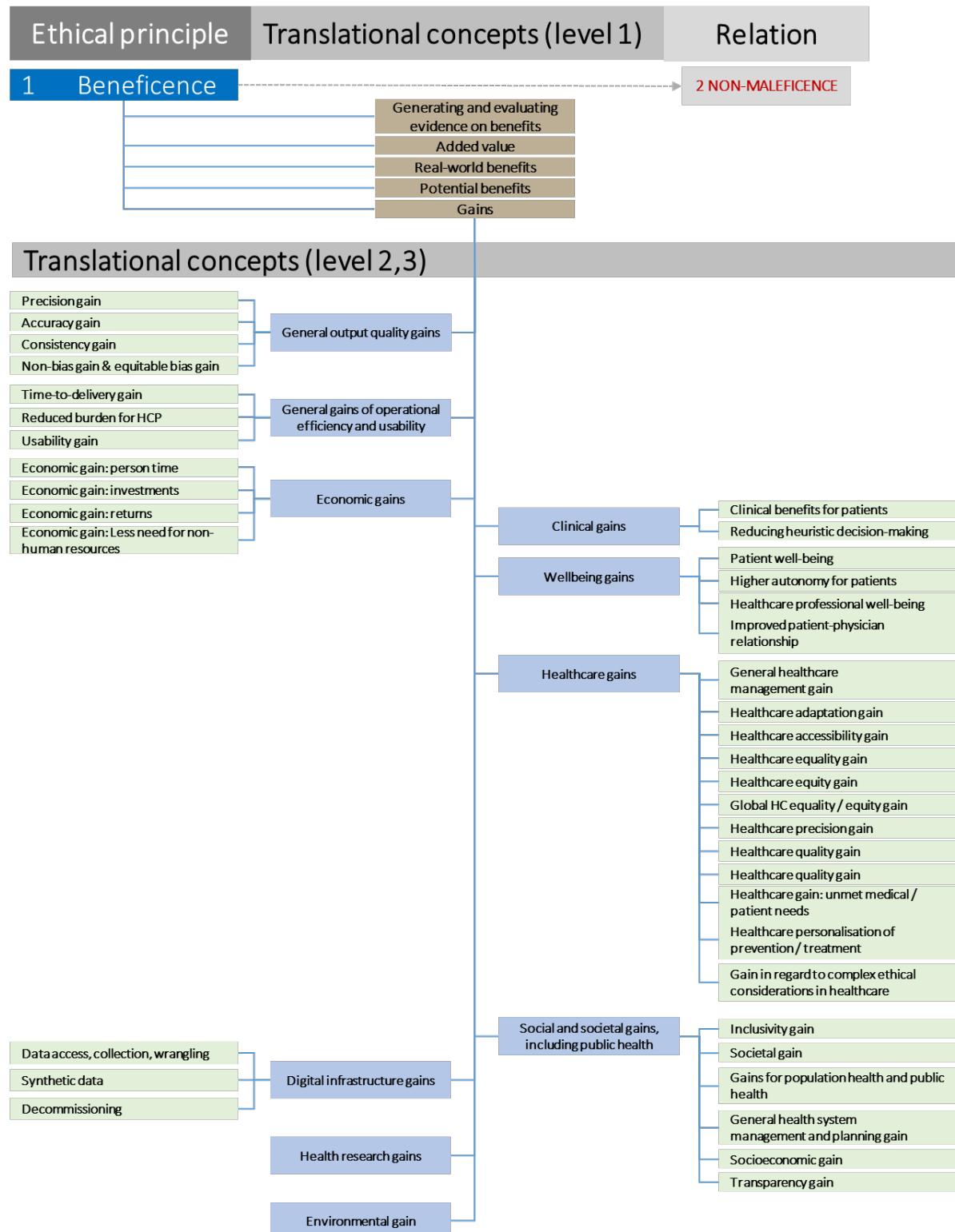
- Bioethics
- AI ethics
- Ethical principles
- Ethical evaluation of AI
- AI principles and ethics guidelines
- Clinical benefit
- Clinical effectiveness

## Term relationship

Related terms:

- TRUST AND TRUSTWORTHINESS

Ontological organisation of the ethical principle of “beneficence” and its translational concepts. Relations to other ethical principles are indicated.



## Beneficence: translational concepts

### Generating and evaluation evidence on benefits

#### Parent term: Beneficence

Evidence on benefits of AI systems, in particular those used in healthcare, is generated throughout the AI evidence pathway. Initially, at the design and concept phase, evidence will not be available and potential benefits will need to be estimated based on preliminary or indirect information. Estimates of potential benefits may draw also on assumptions and judgements of similarity about design characteristics that are considered pivotal for given benefits. This may involve the use of information on other systems available in relevant scientific or clinical literature.

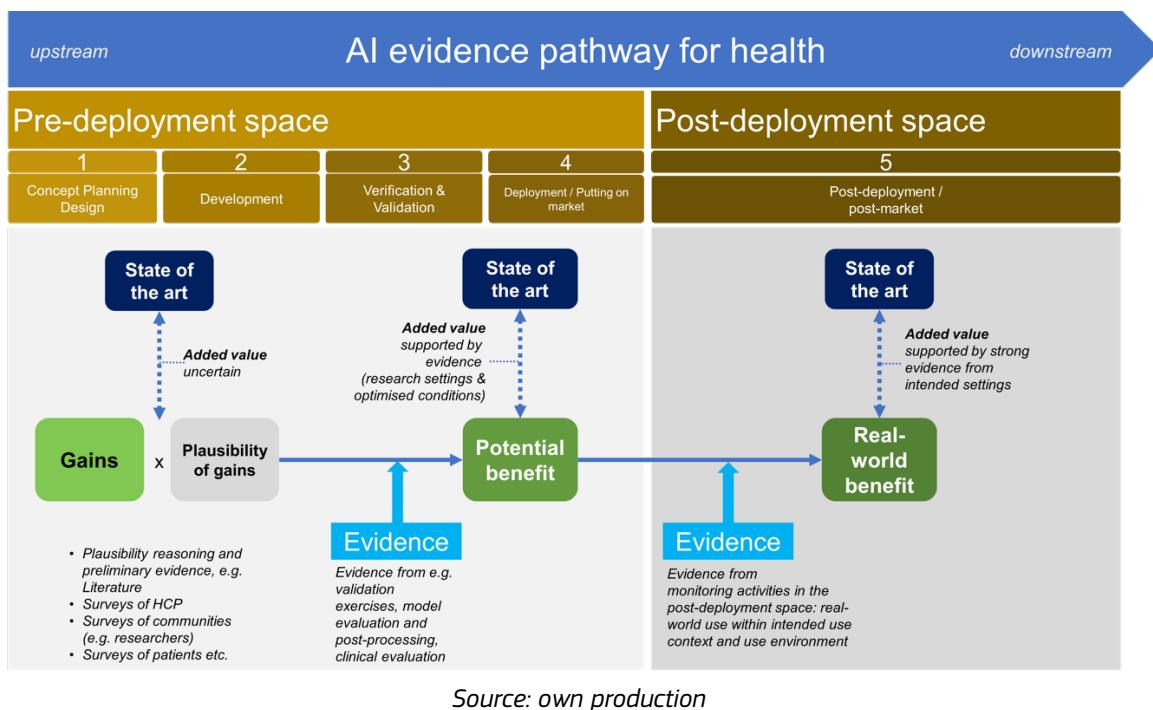
Such preliminary potential evidence should later be **substantiated** by robust direct evidence concerning the specific AI system in question. Typically, initial evidence will be generated under restricted, optimised conditions in research settings (see → validation, → model validation, → model evaluation) in the pre-deployment space, followed eventually by evidence derived from using the AI system in the post-deployment space under non-optimised real-world conditions in a specific → use context and → use environment, providing additional insights on → generalisability, potential → biases and other pitfalls that may need to be addressed in a further iteration of AI system improvement (→ life cycle of AI in health).

Such evidence may be generated by → post-deployment monitoring or, for healthcare uses, during the ongoing process of → clinical evaluation which includes in the post-market space → post-market surveillance, market surveillance, corrective action as well as → post-market clinical follow-up. Sufficient evidence on benefits plays a pivotal role for regulatory purposes (e.g. benefit-risk assessment) and health technology assessment (HTA).

To characterise the trajectory of evidence from the early stages of the life cycle to the post-deployment stage (see **Figure 11**) we use four concepts:

- **added value**: value provided by benefits that go beyond the benefits of a benchmark, e.g. the current state of the art
- **gains**: anticipated benefit during the design and development stage supported by little evidence (e.g. from literature)
- **potential benefits**: benefits supported by evidence predominantly from research settings and obtained under optimised conditions.
- **real-world benefits**: benefits supported by strong evidence from real-world settings (e.g. intended settings or alternative → use environments where the AI system is successfully used)
- Estimations of the plausibility of → gains for an initial estimation of → potential benefits may draw on various information sources, e.g. scientific/clinical literature, surveys of users (e.g. researchers, healthcare professionals, patients; → users of AI in the health domain).
- Evidence on → potential benefits will typically be generated in research settings (e.g. clinical investigation, AI system validation) and typically under optimised conditions (e.g. → model evaluation, involving independent patient groups under controlled conditions).
- Evidence on → real-world benefits will be derived from → post-deployment monitoring activities and, for AI systems used for healthcare, continuous → clinical evaluation and post-market surveillance (→ post-market surveillance, market surveillance, corrective action).

**Figure 11.** Schematic depiction of the relationship of the concepts 'gains', 'added value', 'potential benefit' and 'real-world benefit' under the ethical principle of BENEFICENCE and along the AI evidence pathway.



## Added value

### Parent term: Beneficence

Value or usefulness added by the **AI system's benefits** when compared to the current commonly used practice or to the generally acknowledged state of the art (SOTA): see also → **model actionability** (for AI systems in healthcare).

Importantly, in the *healthcare context*, SOTA does not refer to the most advanced or innovative technology, but to what would be provided to a patient through standard care (see definition of SOTA below).

During the design and development phase, potential added value can be estimated based on literature or limited studies (e.g. in the context of → **validation exercises**, e.g. → **usability validation**). Proven added value is here understood as the sum of → **real-world benefits** for which there is robust evidence, e.g. from → **clinical evaluation** (including during the post-market stage), from user surveys, patient-reported outcomes etc.

### Box 3. State of the art

We refer to the definition of SOTA as provided by IMDRF (2018) (modified from ISO/IEC Guide 2:2004).

*"State of the Art: Developed stage of technical capability at a given time as regards products, processes and services, based on the relevant consolidated findings of science, technology and experience."*

*NOTE1: The state of the art embodies what is currently and generally accepted as good practice in technology and medicine. The state of the art does not necessarily imply the most technologically advanced solution. The state of the art described here is sometimes referred to as the "generally acknowledged state of the art".*

N.B. SOTA plays also a role for safety of medical device technology (→ **AI safety**; → **NON-MALEFICENCE**). The SOTA should be taken into account when considering the benefit-risk ratio of medical device technology (e.g. EU's medical devices Regulation, Annex I, Chapter I, point 1; [EU, 2017](#)).

## Real-world benefits

### **Parent term:** Beneficence

Real-world benefits are real-world positive impacts of the AI system, when used as intended and, where appropriate, in a given defined → use context and/or → use environment. Real-world benefits provide specific, tangible and, ideally, quantifiable → added value as compared to a benchmark (usually the state of the art, SOTA; see → added value).

Consequently, already at early design stages, both → use context and → use environments as well as associate requirements should be considered. This includes

- Required competency of users/operators,
- Competences and needs regarding value chain enablers:
  - IT infrastructure needs and variations of available infrastructure (hospitals or at the patient's home in case of home care)
  - enabling technologies required for the successful, safe and secure operation of the AI system
- inter-operability of products (e.g. AI system with a cloud platform for data storage)
- cybersecurity requirements (that can be reasonably expected at the → use environment).

Investigating ways to ensure best possible integration of the AI system into the real-world → use environment is a key consideration for ensuring that → potential benefits can translate into real-world benefits in the post-deployment stage, once the system is routinely used. This may include uses under conditions that differ from those studied during relevant → verification, → validation (e.g. → model validation, → clinical validation, usability validation) and evaluation exercises (notably those under → clinical evaluation).

See also NON-MALEFICENCE for potential risks. Consideration of risks is critical for defining conditions and requirements needed to ensure real-world benefits. In particular, see:

- risks relating to technical integration and interoperability issues
- risks related to transparency (e.g. unclear instructions for use, unclear training requirements)

## Potential benefits

### **Parent term:** Beneficence

A potential benefit is the product of a given intended → gain and the associated *plausibility* to leverage that specific gain under real-world conditions. A potential benefit is not yet a → real-world benefit, since various obstacles may hinder a specific potential benefit to come to full fruition under realistic use conditions (e.g. real-world → use contexts and → use environments) in intended settings.

Potential benefits may be based on various elements, e.g. the intended purpose, simplification of specific processes, higher accuracy, greater ease of use, reduced bias, absence of biased expert judgments etc. (see translational concepts of gains).

## Gains

### **Parent term:** Beneficence

A 'gain' is a projected, intended, anticipated benefit of an AI system at the pre-deployment stage i.e. before the AI system is used under real-world conditions. The product of a gain and the plausibility of the gain to materialize is understood here as → potential benefit (supported by preliminary evidence).

### General output quality gains

#### **Parent terms:** Beneficence - Gains

The AI system enhances quality-related aspects of a given workflow, process or use case. Specific gains include those outlined below.

#### Precision gain

##### **Parent terms:** BENEFICENCE – Gains - Output quality gains

The AI system enables output that is more precise than when obtained by current systems. This includes for instance enhanced predictive → precision (positive predictive value) or higher precision of specific processes and procedures (e.g. surgical procedures), higher precision of a treatment (e.g. better image segmentation allowing precise dosimetry and hence less side-effects), albeit not yet on a health system level (see for comparison → healthcare precision gains).

#### Accuracy gain

##### **Parent terms:** BENEFICENCE – Gains - Output quality gains

The AI system enables output that is more accurate than when obtained by systems deemed current situational or global state of the art. This concerns mainly predictive → accuracy.

#### Consistency gain

##### **Parent terms:** BENEFICENCE – Gains - Output quality gains

The AI system supports a higher degree consistency of the results of a specific process or procedure (less variability of outcomes). This could concern surgical procedures, diagnostic test workflows, clinical workflows, research processes (e.g. knowledge management and mining (e.g. of big health data), pattern recognition, predictions).

#### Non-bias gain & equitable bias gains

##### **Parent terms:** BENEFICENCE – Gains - Output quality gains

###### *Undesired bias*

The AI system reduces undesired and discriminatory → bias as compared to the situational or global state of the art (e.g. in screening, diagnosis, prognosis approaches, patient recruitment for clinical investigations/trials etc.). See also → digital infrastructure gains.

###### *Desired, equitable bias*

Specific → bias (equitable bias; see → health equality & health equity) that ensures that a given AI systems is targeted and performs for vulnerable groups and/or patient with rare diseases / conditions (see also → universal versus targeted design).

## General gains of operational efficiency and usability

**Parent terms:** Beneficence - Gains

The AI system supports more efficient operations of the given use environment, leading to positive spill-over effects for the organisation, employees and stakeholders. Specific gains include those outlined below.

### Time-to-delivery gain

**Parent terms:** BENEFICENCE – Gains - Operational gains

The desired outcome of a specific process or procedure can be delivered faster as compared to the current situation.

### Reduced burden for HCP

**Parent terms:** BENEFICENCE – Gains - Operational gains

The health care professional's work burden is reduced, allowing for instance to focus more on patient interaction and discussions with patients.

### Usability gain

**Parent terms:** BENEFICENCE – Gains - Operational gains

The AI system shows comparatively higher → usability (e.g. less training requirements, ease of use, clarity of user interface) when compared to the state-of-art or relevant benchmarks. This may also reduce error rates in specific workflows, clinical practice (→ clinical practice protocol, → clinical practice guideline) or → clinical pathways.

## Economic gains

**Parent terms:** Beneficence - Gains

The AI system allows to achieve the output with less resources, less cost, less investments.

Specific gains include those outlined below.

### Economic gain - person time

**Parent terms:** BENEFICENCE – Gains - Economic gains

The AI system may allow to achieve the same output with less person-time, when involving the same number of persons or with less persons as compared to the situation without AI system use. This gain may however lead to negative socioeconomic consequences, e.g. displacement of human workers by the AI system.

### Economic gain: investments

**Parent terms:** BENEFICENCE – Gains - Economic gains

The AI system may allow to achieve the same output with less investments required, resulting in budget benefits. This does not include person time gains are captured above.

### Economic gain: returns

**Parent terms:** BENEFICENCE – Gains - Economic gains

The AI system's output may lead to additional returns in comparison to the previous.

### Economic gain: less need for non-human resources

**Parent terms:** BENEFICENCE – Gains - Economic gains

More effective usage of non-human resources, e.g. due to better planning and timely procurement of material.

## Clinical gains

### **Parent terms:** BENEFICENCE – Gains

For AI systems used in healthcare: the AI system provides clinical gains to patients.

## Clinical benefits for patient

### **Parent terms:** BENEFICENCE – Gains – Clinical gains

The AI solutions may bring → **clinical benefits for patients** that were previously not attainable.

Definitions of clinical benefit:

- Article 2 (53) of the EU MDR defines clinical as the positive impact of a device on the health of an individual, expressed in the terms of a meaningful, measurable, patient-relevant clinical outcome(s), including outcome(s) related to diagnosis, or a positive impact on patient management or public health; whereas
- Article 2 (37) of the EU IVDR defines clinical benefit as the positive impact of a device related to its function, such as that of screening, monitoring, diagnosis or aid to diagnosis of patients, or a positive impact on patient management or public health.<sup>6</sup>

Source: EU 2017/745 (MDR), Article 2 (53); EU 2017/746 (IVDR), Article 2 (37) and IVDR recital (64)

For clinical benefits for patients based on personalised medicine, see → **healthcare gain – personalisation of prevention/treatment**.

## Reducing heuristic decision-making

### **Parent terms:** BENEFICENCE – Gains – Clinical gains

The AI system reduces heuristic decision making (→ **heuristics**), e.g. by allowing for a more complete evaluation of clinically relevant information and/or their better integration, reducing possible human → **bias** and improving the quality of diagnostic decisions, treatment decisions and overall clinical care.

## Well-being gains

### **Parent terms:** Beneficence – Gains

The AI system provides potential gains in regard to the well-being of persons.

## Patient well-being

### **Parent terms:** BENEFICENCE – Gains – Well-being gains

The well-being of the patient is improved, e.g. due to less burdensome procedures, reduced need for diagnostic testing, more precise predictions, reduced need to visit the clinic, possibilities for home care and patient empowerment to self-manage (where desired).

Higher degree of autonomy / independence of patients with chronic diseases and/or debilitating conditions

### **Parent terms:** BENEFICENCE – Gains – Well-being gains

Patients with chronic diseases and/or debilitating conditions gain a higher degree of independence from healthcare structures, allowing them to lead a freer life with more autonomy.

## Healthcare professional well-being

### **Parent terms:** BENEFICENCE – Gains – Well-being gains

The health care professional's well-being is enhanced (e.g. due to → **operational gains**), leading to a better work experience, more attractive working conditions, reduced stress and possibility to focus more on patient interaction.

## Improved patient-physician relationship

### **Parent terms:** BENEFICENCE – Gains – Well-being gains

The relationship between patient and physician is improved, e.g. due to enhanced clarity of diagnostic predictions, enhanced interaction time etc.

## Healthcare gains

**Parent terms:** Beneficence - Gains

The AI system provides potential gains in regard to healthcare delivery. Specific gains include those outlined below.

### General healthcare management gains

**Parent terms:** BENEFIENCE – Gains - Healthcare gains

The AI solution supports healthcare administration at healthcare settings, e.g. patient workflow planning, scheduling, patient communications, workflow planning of healthcare professionals, billing, report writing (e.g. based on notes or recorded conversations), integration of new information into electronic patient records, medical coding etc.

### Healthcare adaptation gains

**Parent terms:** BENEFIENCE – Gains - Healthcare gains

The AI system supports the continuous and dynamic adaptation of healthcare systems to new innovative solutions and their integration into healthcare practices, clinical workflows.

### Healthcare accessibility gains

**Parent terms:** BENEFIENCE – Gains - Healthcare gains

The AI system may enhance accessibility to healthcare for individuals, at a local or community level or within a country or region or socioeconomic bloc (e.g. EU). See also → healthcare equality gains and → healthcare equity gains.

### Healthcare equality gains

**Parent terms:** BENEFIENCE – Gains - Healthcare gains

The AI system may enhance healthcare equality for all groups and strata of a given population or community.

### Healthcare equity gains

**Parent terms:** BENEFIENCE – Gains - Healthcare gains

The AI system may enhance the equity of healthcare provided to individuals, independent of race, ethnicity, ability, age, gender etc.

### Global healthcare equality & equity gain(s)

**Parent terms:** BENEFIENCE – Gains - Healthcare gains

The use of the AI system enhances healthcare equality and equity on a global level, e.g. by providing an affordable and/or mobile solution previously unattainable, allowing to manage the health of remote communities with little or no access to clinics or to improve regional or national healthcare systems.

### Healthcare precision gain

**Parent terms:** BENEFIENCE – Gains - Healthcare gains

The AI system may support a more precise treatment of patients at a larger scale of delivery (i.e. at health system level). This includes interventions tailored to individual needs, genetic background, comorbidities and health susceptibilities / risks or tailoring to specific communities. See also → precision gains; → healthcare gain – personalisation of prevention/treatment

### Healthcare quality gain

**Parent terms:** BENEFIENCE – Gains - Healthcare gains

The AI system improves the quality of healthcare on a larger scale.

### Healthcare efficiency gain

**Parent terms:** BENEFIENCE – Gains - Healthcare gains

The AI system improves the efficiency of healthcare processes on a larger scale.

## Healthcare gain - unmet medical or patient needs

**Parent terms:** BENEFIENCE – Gains - Healthcare gains

The AI system allows to address health problems that previously could not be (well) managed or addressed (diagnostics, treatment, prevention) or meets patient needs that were previously unattainable. N.B. For a critical discussion of the term “unmet medical need”, please see: Vreman RA et al. (2019) Unmet Medical Need: An Introduction to Definitions and Stakeholder Perceptions, Value in Health, 22(11): 1275-1282. Online: <https://doi.org/10.1016/j.jval.2019.07.007>.

## Healthcare gain - personalisation of prevention/treatment

**Parent terms:** BENEFIENCE – Gains - Healthcare gains

The AI solution allows enhanced tailoring of diagnosis, prevention or treatment to the specific requirements of specific groups of individuals or even individual patients (→ personalised medicine and precision medicine).

## Gain in regard to complex ethical considerations in healthcare

**Parent terms:** BENEFIENCE – Gains - Social, societal and environmental gains

The AI system provides output that supports complex ethical considerations (“ethical conundrums”; Arnold, 2021; Wallach et al., 2008) in the context of clinical decision making, clinical practice and/or health system planning and management.

## Social and societal gains, including public health

**Parent terms:** Beneficence – Gains

The AI system provides potential gains in regard to social, societal and environmental dimensions, including inclusivity (non-discrimination), transparency and ethics. Specific gains include those outlined below.

## Inclusivity gain

**Parent terms:** BENEFIENCE – Gains - Social, societal and environmental gains

Use of the AI system leads to or supports a more inclusive approach concerning minorities or specific (vulnerable) patient populations as compared to the situational or global state of the art.

## Societal gain

**Parent terms:** BENEFIENCE – Gains - Social, societal and environmental gains

The AI system has positive effects on society (i.e. a large number of people, part of the population or the entire population) and is not-confined to a specific (user)community. Examples include “democratisation” of medicine, enhanced awareness of health risk factors supporting healthy lifestyle choices within a society, enhanced freedom (to choose), patient empowerment, reduced burden of disease for the entire society etc. (see also → gains for population health and public health).

## Gains for population / public health

**Parent terms:** Beneficence – Gains

The AI system provides potential gains for measures aimed at improving population health through, for example:

- screening, diagnosis, treatment, rehabilitation approaches, reducing health inequities and inequalities within and across the population.
- health promotion measures
- disease prevention, prediction-based public health surveillance, monitoring of disease outbreaks and disease spreads, preparedness of cross-border health threats and emergencies
- improved disease outbreak responses and improved targeted public health and social measures (PHSM) with the aim of balancing health risks versus socioeconomic impacts (WHO,

[2024d](#)), including robust methodologies for gathering evidence on the effectiveness of PHSM measures ([Fadlallah et al., 2024](#)).

## General health system management and planning gain

**Parent terms:** BENEFIENCE – Gains - Social, societal and environmental gains

The AI system contributes to the improved management and planning of health systems on various levels (national, regional or local) in view of for instance

- improving the impact of health policies (including prevention and screening) and the effectiveness, equity and equality of healthcare delivery – in particular between urban and rural or otherwise underserved areas and communities.
- improved supply chain management, procurement, effective communication strategies, forward-looking healthcare workforce planning,
- joint systems for patient scheduling, joint databases of electronic health records supporting primary and secondary use of health data etc.
- improved common approaches for cybersecurity of health systems and healthcare settings, including cyber incident characterisation, cyber resilience and health data protection ([Reina & Griesinger, 2024a, b](#)).

## Socioeconomic gain

**Parent terms:** BENEFIENCE – Gains - Social, societal and environmental gains

The AI system has positive socio-economic effects, e.g. job creation, shift from repetitive to more education-intensive jobs, opportunities for start-ups etc.

## Transparency gain

**Parent terms:** BENEFIENCE – Gains - Social, societal and environmental gains

The AI system improves the transparency of a specific process, or procedures as compared to the situational or global state of the art.

## Digital infrastructure gains

**Parent terms:** Beneficence - Gains

The AI system provides potential gains in regard to digital infrastructure needed and/or supporting the development, deployment, use, monitoring and → decommissioning of AI systems. Specific gains include those outlined below.

## Data access or collection, wrangling

**Parent terms:** BENEFIENCE – Gains - Digital infrastructure gains

The AI system supports the identification of lawfully accessible high-quality health data (for training of other AI systems), access and/or collection to these data, management of privacy aspects and/or data compilation, data labelling and data wrangling activities (e.g. categorizing data, harmonising data fields, data set creation). This includes addressing → bias (see also → non-bias gains & equitable bias gains).

## Synthetic data

**Parent terms:** BENEFIENCE – Gains - Digital infrastructure gains

The AI solution contributes to the generation and availability of high-quality synthetic health data (→ synthetic data) for the development and training of AI systems for health and medicine.

## Decommissioning

**Parent terms:** BENEFIENCE – Gains - Digital infrastructure gains

The AI system supports decommissioning by enabling more effective and simpler decommissioning of the systems and data.

## Health research gains

**Parent terms:** Beneficence – Gains

The AI system provides potential gains in regard to health research. For example: drug design, development, repurposing, clinical study/trial design, clinical data evaluation, exploitation of big health data (e.g. aggregated electronic health records), conduct of in silico trials, virtual human twins etc.

## Environmental gain

**Parent terms:** BENEFIENCE – Gains - Social, societal and environmental gains

The use of the AI system has positive effects on the environment and planetary health, e.g. less use of resources (energy, consumables, water, CO<sub>2</sub> footprint & climate impact, reduced use of antibiotics, reduced pharmaceutical use resulting in less degradation-based substances in waste water etc.).

This includes also AI systems that, on a comparative level, require less resources (e.g. electricity) and have hence a lower environmental footprint than other AI solutions.

## A.2 Non-maleficence

### Concept description

#### **Non-maleficence in the context of AI systems**

We use NON-MALEFICENCE in the context of AI systems for health in a broader sense than in → bioethics (see below). We understand non-maleficence as the obligation of relevant actors involved in the development, deployment, assessment, making available of and in the use of AI systems, to avoid anything that might cause harm to or put unnecessarily at risk

- patients, users and other persons
- groups, communities and society.

Thus, non-maleficence has a broader scope than safety (→ AI safety), which concerns safety of patients, users and other persons (see also [WHO, 2003; GHTF, SG1-N020R5M, 2005; IMDRF, 2018; EU, 2017a](#)).

The obligation of non-maleficence requires that there is

- a) sufficient **awareness of various sources of risks associated with AI** which will support, in the context of healthcare, benefit-risk assessments (e.g. under → AI risk management processes) during → clinical evaluation (see also → BENEFICENCE).
- b) sufficient **consideration of wider (harmful) impacts of algorithms** on groups, societies, fundamental rights and other ethical aspects are sufficiently considered (see also → FAIRNESS; → AI impact assessment; → fundamental rights and algorithm impact assessment, → ethical evaluation of AI).

The translational concepts of NON-MALEFICENCE reflect both elements:

- a) sources of AI-associated risks: most can be related to other ethical principles and
- b) impact assessments.

#### **Non-maleficence in bioethics and clinical ethics**

Non-maleficence is one of the four principles of → bioethics and its subfield of clinical ethics. In that context non-maleficence primarily relates to harm and risks for patients. For instance: „*Nonmaleficence means to do no harm by refraining from providing ineffective or harmful treatments as determined by whether the benefits of treatment outweigh the burdens*“ ([Jankowski, 2014](#)).

Thus, NON-MALEFICENCE in healthcare incorporates considerations of a positive benefit-risk ratio (→ AI risk management; see for instance EU MDR Annex I, Chapter I; [EU, 2017a](#)). The ethical foundation of non-maleficence is the human right to life and right to the integrity of the person, i.e. respect for physical and mental integrity (see for instance EU charter of human rights, Articles 2 and 3; [EU, 2000](#)).

#### **Sources of AI-associated risks for persons, groups, community and society**

Risks related to ethical principles:

- Risks related to → DIGNITY, FREEDOM AND AUTONOMY, e.g. informed consent to medical treatments or risks concerning consequential decision making based on AI (see [Calo R, 2018](#)).
- Risks related to → PRIVACY PROTECTION, e.g. privacy of personal information and health information. Risks include data breaches, data theft (including both training data and post-deployment data, e.g. clinical data produced in the post-deployment space by an AI-enabled diagnostic tool).
- Risks related to → TRANSPARENCY, e.g. lack of → interpretability and explainability of input-output behaviour affecting the right for informed consent to medical treatments.
- Risks related to → FAIRNESS, including

- Risks due to → bias, non-inclusion, discrimination, e.g. AI tools that are inaccurate for specific patient groups.
- Risks for specific patient groups concerning access to healthcare, equal treatment and equitable outcomes.
- Risks related to → RESPONSIBILITY), e.g. due to lack of → accountability structures that may delay necessary corrective actions and associated communication or risks related to human agency and oversight (→ ensuring human agency and oversight).
- Risks related to → SOLIDARITY, e.g. socioeconomic risks, e.g. workplace impacts (e.g. job losses, devaluing of jobs, skills, education), leaving people (unintentionally) behind.
- Risks related to → SUSTAINABILITY (e.g. environmental risks).

In addition, there are risks related to:

- insufficient robustness
- technical integration and interoperability
- integration of AI systems into workflows (e.g. clinical pathway)
- pre-deployment evidence gaps (e.g. lack of sufficient validation)
- insufficient post-deployment monitoring or post-market surveillance.

#### **Explanatory note**

##### ***Non-maleficence, AI safety and AI risk management***

NON-MALEFICENCE is closely linked to → AI safety and → AI risk management (**Figure 12** below).

Briefly the relationship between non-maleficence, AI safety and AI risk management can be summarized as follows:

- Non-maleficence is the obligation of actors to avoid possible harm and, hence, sufficiently control risks for patients, users and other persons to ensure → AI safety and to understand and mitigate wider impacts, e.g. on groups, communities or society.
- → AI safety is the state of being protected from harm associated AI-specific risks. AI systems should be developed, deployed and used in a way that does not compromise the safety of patients, users and other persons. AI safety depends on non-maleficence. AI systems, in particular in healthcare, may require additional safety considerations not linked to AI.
- Safety is typically addressed in the context of structured risk management processes (see for instance EU's MDR, Annex I, Chapter I, point 3 ([EU, 2017a](#)). Consequently, → AI safety will be tackled by → AI risk management (e.g. see EU's AI Act, Article 9 ([EU, 2024a](#))). Notably, risk management includes also non-safety risks such as risks for property, an organisation or the natural environment. Further, → AI risk management processes may also address the wider impacts on groups, communities and societies (see also: → AI impact assessment; → fundamental rights and algorithm impact assessment, → ethical evaluation of AI).

#### ***Realising non-maleficence through risk management and other processes***

Risk management (→ AI risk management) implies identifying, framing, controlling and managing harms and risks, i.e. lowering the probability of harm to occur. Risk management is typically approached through dedicated frameworks, processes or organisational systems (e.g. [NIST, 2023](#)).

In addition to standardized processes of risk management that may have a focus on technologies, additional approaches may be required to address risks related to, for instance, job impacts, deskilling, a trustful patient-physician relationship (see **Box 4** under → generating and evaluating evidence on risks).

### **Non-maleficence requires collaboration along the life cycle stages and across the value chain**

Due to the complexities of AI technologies in regard to the → **life cycle of AI in health** and the → **value chain of AI**, responsibilities may be highly distributed among various organisations, companies and actors. This can complicate relevant corrective actions in case of problems, failures or malfunctions. Collaboration along the → **AI evidence pathway for health** is therefore required to ensure that relevant evidence on risks, harms as well as their adequate control is available.

### **Maximising clinical benefits and minimising harm: AI safety, performance and equity**

For healthcare products, a positive benefit-ratio is a key consideration for acceptability of risks (→ **AI risk management**; → **risks**). In a broader sense, AI systems in healthcare should be developed and deployed so that clinical benefits are maximized and harms minimized.

This requires taking *performance* of an AI system into consideration: unsatisfactory performance or unacceptably high variations of performance (e.g. depending on → **use environment**) may constitute a risk for patients and hence a health safety concern (apart from being a concern regarding → **health equality & health equity**).

The IMDRF's guidance on "Essential principles of safety and performance of medical devices and IVD medical devices" (IMDRF, 2018) outlines general aspects of design and manufacturing of medical devices and IVD medical devices to ensure that they are safe and perform as intended. These are relevant for → **AI-enabled medical device software**.

For risk-benefit considerations and thus the determination of acceptable risks, there should be evidence on the following three aspects of an AI system (see also Char et al. 2020):

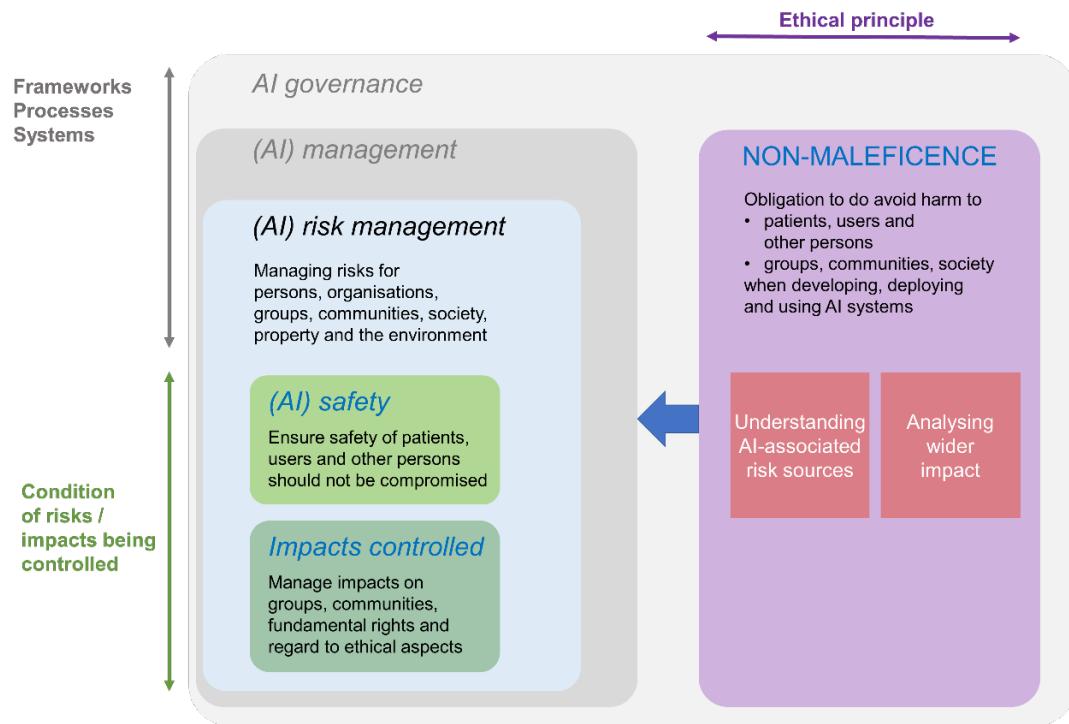
- *AI safety* for patients, users and other persons:
  - Evidence on the minimisation of harm and the probability of harm to occur (=risk) when the AI system is used in line with the intended purpose and as foreseen by the manufacturer / developer.
  - Evidence on residual risks and a rationale why these are deemed acceptable when considering evidence on benefits.
  - AI safety requires conscientious approaches to non-maleficence, e.g. in the context of → **AI risk management** processes.
- *Clinical performance and effectiveness*:
  - Evidence that the application helps addressing or solving the health problem it was designed for at a reasonable cost, in particular in regard to patient health (e.g. cost of false classifications)
- *Equity*:
  - Evidence that benefits of the AI solutions are shared by all and accessible to all.

### **Term relationship**

Related terms:

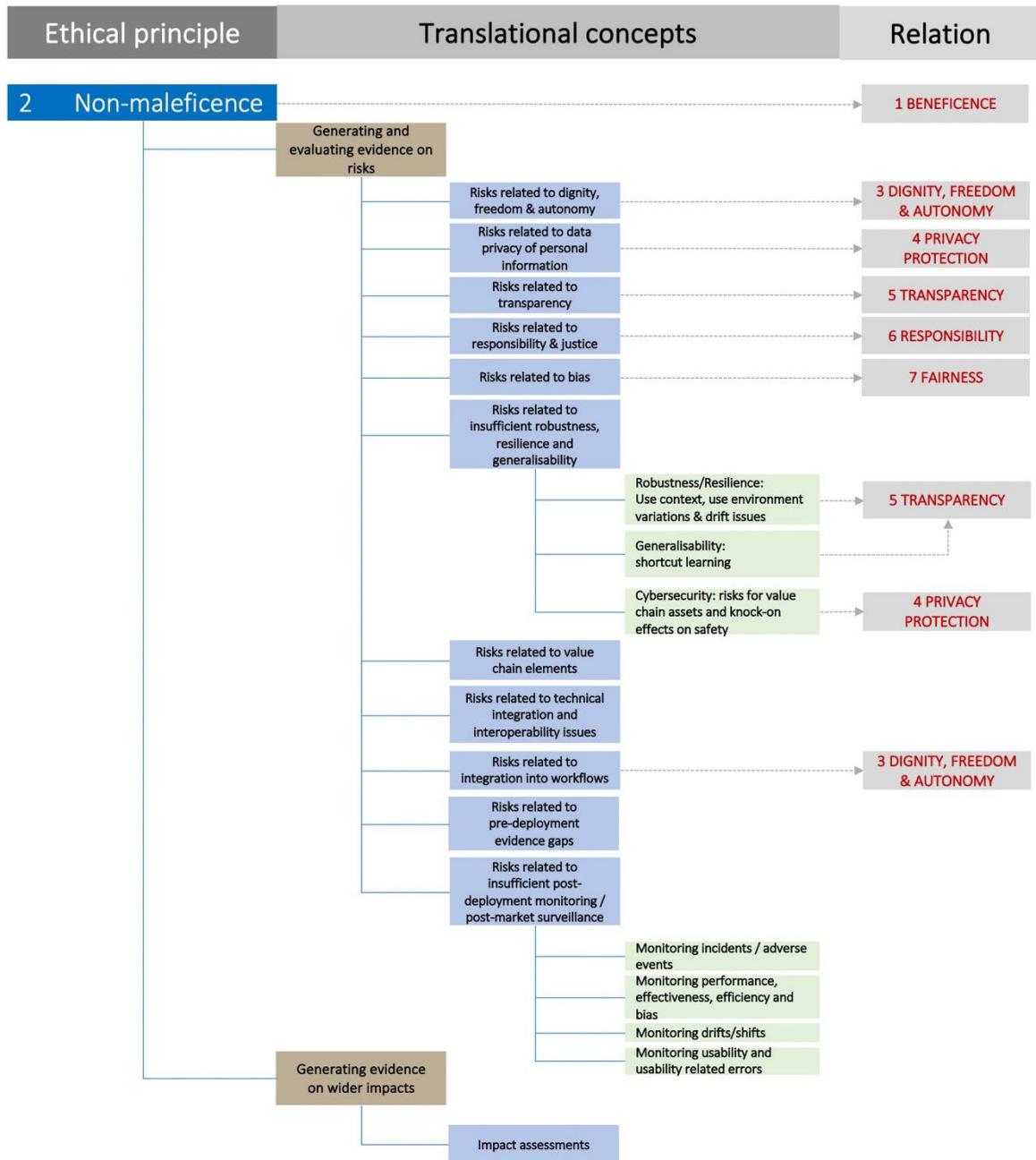
- **AI safety**
- **AI risk management**
- **Risk**
- **AI management**
- **AI governance**

**Figure 12.** Schematic depiction of the relationship of the concepts of NON-MALEFICENCE, → AI safety, → AI risk management, → AI management and → AI governance. N.B. The diagram is highly schematic. AI management, AI risk management and AI governance refer to **frameworks, processes or systems**. AI safety and controlled impacts to the **condition** of risks and impacts being adequately controlled. NON-MALEFICENCE to the **ethical obligation** to avoid harm, necessitating an **understanding of AI-associated risks and wider impact**. We understand NON-MALEFICENCE as an element of AI governance that feeds into AI management and in particular into AI risk management.



Source: own production

Ontological organisation of the ethical principle of “non-maleficence” and its translational concepts. Relations to other ethical principles are indicated.



## Non-maleficence: translational concepts

### Generating and evaluating evidence on risks

#### Parent term: Non-maleficence

Evidence on possible risks as well as evidence concerning continuous proactive → AI risk management activities needs to be generated in order to ensure → AI safety and to uphold, in particular in healthcare, the principle of → NON-MALEFICENCE in regard to patients, users and other persons.

This requires a proactive approach of looking out, framing, describing and mitigating risks. Such an approach requires a **good understanding of potential harms, risks and pitfalls**. While each AI system is unique, risks associated with development and use of AI systems are not. This ontology is intended to support a proactive and risk-informed approach. Risks need to be evaluated

- Throughout the → life cycle of AI in health and
- Across the → value chain of AI (→ risks related to value chain elements), e.g. related to the use of purchased → development data or pretrained models
- Typically, many actors are involved in design, development, deployment, evaluation and use of AI systems. Thus, various → AI actors and communities need to collaborate to ensure that risks are identified wherever they arise.

Risk management typically focuses on processes of risk identification and control. With the concept of the → AI evidence pathway for health, we emphasise the importance of evidence on how risks were identified, how they are minimized or mitigated and how residual risks have been deemed acceptable in regard to the benefits (→ BENEFICENCE), taking into account also the → intended use of the AI system as well as the → use context and → use environment.

The following translational concepts, lay out risks in relation to ethical principles, the value chain and implementation of AI systems.

#### Box 4. Other risks not related to non-maleficence and AI safety:

- There may be other potential negative effects of AI use on healthcare professionals that are not safety risks, i.e. they do not pose an *immediate* danger to life or physical or mental integrity. These are considered under → DIGNITY, FREEDOM AND AUTONOMY (e.g. deskilling; patient-physician relationship) and → SOLIDARITY (e.g. job impacts, skills and training).
- Organisational risks are only considered in this ontology where they are connected to safety: e.g. risks related to value chain elements may carry both risks for property of companies and also for patients; only the latter are considered here. Such risks are addressed usually through → AI risk management.
- Risks for the environment are implicated under → SUSTAINABILITY.

### Risks related to dignity, freedom and autonomy

#### Parent terms: Non-maleficence – Generating and evaluation evidence on risks

Depending on their design and the way they are used in clinical practice, AI systems may pose risks for the dignity of patients, affecting their freedom and autonomy regarding a number of topics, detailed under → DIGNITY, FREEDOM AND AUTONOMY.

Briefly:

- There may be risks related to the innovativeness of AI solutions and/or efficiency gains sought by health systems, which may impair the primacy of patient care over scientific or societal interests (→ respecting patient primacy).
- AI will influence the future development of healthcare (→ AI and the development of healthcare), with risks from a variety of perspectives, including the potential to implement practices of 'nudging' (particularly easy to achieve with AI agents), the 'economisation' of health care or adverse behavioural impacts on patients/users due to the wide-spread use of → conversational agents (chatbots).
- A deterioration of the patient-physician relationship represents a major and multi-faceted risk (→ upholding a trustful patient-physician relationship). This includes issues such as → avoiding automation bias, → avoiding automation complacency and the → deskilling of healthcare professionals, due to reliance on automated systems that provide data integration and interpretation (e.g. radiological imaging, patient information for clinical decision making). Importantly, attention to patient-specific needs as well as the required attunement of the healthcare professional to the patient, his/her history and preferences, may deteriorate.
- AI and in particular 'black-box' AI (i.e. systems that are not inherently interpretable and hence unintelligible) pose considerable risks for patient dignity in regard to the free and informed consent of patients (→ ensuring the means for free and informed consent).
- The potential effect of an AI system on the patient's right to know but equally the right not to know about his/her medical condition needs to be considered (→ right to know and right not to know).
- Inadequate design and/or implementation of AI systems (in particular → conversational agents) may pose risks to the right of patients and other users to know when they are interacting with an AI system. This has bearings on trust into healthcare and health systems which may have long-term effects on the effectiveness of public health measures.
- Finally, there are several risks related to → data privacy and more specifically → medical privacy / health privacy.

#### Risks related to data privacy of personal information

##### **Parent terms:** Non-maleficence – Generating and evaluation evidence on risks

These risks relate to inadequate or insufficient measures concerning → PRIVACY PROTECTION, affecting → data privacy of → personal data, including health information (→ medical privacy / health privacy).

There are many entry points of potential risks in this regard along the → life cycle of AI in health and the → value chain of AI.

Privacy of data and personal information are at risk for instance due to inadequate → data protection and → data security processes. For instance, due to inadequate measures

- for making training data non-identifiable (e.g. 'pseudonymisation': techniques for removing, replacing or transforming personal information from data sets) may leave personal data accessible by non-authorised persons, affecting data accountability (→ data governance, data management & data accountability)
- for protecting data in the post-deployment / post-marketing space, including when → decommissioning / retirement AI systems may expose data to unauthorized access, theft or leakage.

Cybersecurity is critical for preserving the integrity and protecting data. We consider cybersecurity a value-preserving enabler of the → value chain of AI, preserving data assets, but also models and AI systems ([Reina & Griesinger, 2024](#)). Cyber attacks may be specifically aimed at data in both the pre- and post-deployment stage.

## Risks related to transparency

### **Parent terms:** Non-maleficence – Generating and evaluation evidence on risks

Insufficiently transparent and clear description of AI systems may pose safety risks due to

- User errors that may be related for instance to insufficient or inadequate descriptions of → intended use, → instructions for use, or → applicability and limitations.
- Similarly, insufficient → information on training requirements and aspects of → usability may lead to inappropriate use of an AI system with potential consequences for reliability, performance. This may imply severe risks for patients, for instance through → foreseeable misuse.
- Lack of clarity regarding the requirements of IT infrastructure, interoperability and relevant value chain elements (e.g. services) may pose significant risks of malfunction affecting also patients (→ evidence on interoperability and value chain elements).
- Lack of → intelligibility of how the AI systems works, how and why it produces specific outputs (as opposed to other ones: counterfactual reasoning) and, in particular lack of → interpretability and explainability of models pose risks. Such situations may lead to misunderstandings of outputs, inadequate use of AI systems and deterioration of the patient-physician relationship: healthcare professionals may in such situations give the AI system the “benefit of the doubt” without sufficiently exercising sufficient oversight or ‘human veto’ (→ ensuring human agency and oversight). Lack of → intelligibility may also frustrate healthcare professionals and degrade their readiness to engage in actions in support of → corrigibility, where necessary.
- Lack of → traceability of processes and decisions relating to the development of an AI system may have serious consequences for patients: it may hinder or at least delay necessary corrective actions in case of malfunction or failure (→ post-market surveillance, market surveillance, corrective action), simply because it will be more difficult to identify at which step along the → algorithm-to-model transition or when deploying or using the AI system, potential errors have occurred. Traceability is however not only an issue for model developers but should be exercised by all actors of the → value chain of AI. Data providers for instance should be able to provide traceable information concerning the → data provenance and their processes for collecting, wrangling data sets (→ data processing / wrangling). This includes cases where collected real-world data were supplemented with → synthetic data.
- While a lack of → transparency of human-AI interaction (e.g. → conversational agents for patient flow management or for emergency triaging) will most likely not pose immediate risks for patients, it may lead to incorrect expectations on patient and user side and, on the mid to long term, degrade confidence in health systems and/or public health.

## Risks related to responsibility

### **Parent terms:** Non-maleficence – Generating and evaluation evidence on risks

There are numerous risks for patients related to insufficient responsibility of communities and organisations that develop, deploy and use AI in the healthcare context.

These include

- Lack of a → quality culture and risk management conducive to identifying, framing, minimising or mitigating risks at various stages of the → life cycle of AI in health and across the → value chain of AI. See all other entries under the level 1 concept of → generating and evaluating evidence on risks (both pre- and post-deployment / post-market).
- Lack of a responsible approach and dedicated structures and workflows for → correcting problems and failures, including necessary communication (e.g. to deployers and users of AI systems).
- Lack of → peer review and community discourse, including

- on → bias, → usability (and possible use errors with safety implications)
- on aspects relating to → DIGNITY, FREEDOM AND AUTONOMY (e.g. → AI and the development of healthcare, → upholding a trustful patient-physician relationship)
- on auditability (→ auditability and auditing), for instance in case there is a lack of → traceability.
- Lack of accountability attribution, accountability structures (see e.g. [Cerna, 2018](#); [NIST, 2023](#); [Mittelstadt, 2021](#)), including for responsibilities that are distributed among various actors or communities along the life cycle and/or value chain (→ accountability structures, attribution of (distributed) responsibilities). This may lead to delays in responding and correcting failures and problems with potential health impacts for patients.
- Risks may stem from insufficient possibilities or discouragement of human oversight (see → ensuring human agency and oversight) and “active veto” by human expert users (see → corrigibility). In particular the possibility of overriding an AI-generated decision may be critical to ensure avoidance of harm, e.g. through implementation of ‘stop buttons’, abortion protocols and adequate user training ([EC HLEG, ALTAI, 2020](#)). Robotic surgery is an example in healthcare where a lack of human oversight and immediate corrective or abortive action might lead to severe health impacts.
- Lack of implementing sufficient human oversight may in addition abet automation complacency and bias (see → avoiding automation bias and → avoiding automation complacency).

## Risks related to bias

### **Parent terms:** Non-maleficence – Generating and evaluation evidence on risks

Bias is the property of systematic errors being made by systems or people (→ bias). Bias of AI systems and people using these carries significant risks for patients (see → bias, heuristics, drift & shift). Notably, discriminatory bias is not only a ‘fairness issue’ (→ FAIRNESS, → avoiding discrimination & discriminatory bias), but also an issue of → AI safety and hence non-maleficence: if an AI system consistently performs worse for patients of a specific group with sensitive → attributes these patients may be misdiagnosed or receive inferior clinical care. During → machine learning (ML), bias may enter at various stages of the → algorithm-to-model transition, e.g. due to

- incorrect or outdated basic scientific / clinical assumptions (→ conceptual relevance, → valid clinical association / scientific validity)
- data collection and processing decisions leading to insufficient, inadequate data that are biased
- data gaps that are filled in with biased → synthetic data
- incorrect data labelling and or use of health-irrelevant features
- model selection decisions involving intrinsic bias in (pretrained) models
- inadequate techniques for addressing bias in regard to fairness, non-discrimination
- gaps in regard to interpretability and explainability, rendering models more prone to → avoiding automation bias and → avoiding automation complacency)
- inadequate → evaluation metrics

However, not all bias is necessarily harmful. There may be, for justifiable reasons, residual and unavoidable bias (e.g. due to limitations of data sets). Moreover, → intrinsic incompatibilities or ‘trade-offs’ may also be considered bias. These are unavoidable in case of predictive AI systems. There may be also situations where bias is deliberately incorporated into an AI system, e.g. as ‘equitable bias’ (see → bias) in situations of designs being specifically targeted at specific (vulnerable) groups which cannot be adequately covered for diagnosis or clinical decision-making by designs of more universal applicability (→ universal versus targeted design; see → FAIRNESS).

Thus, dealing with and adequately tracing and describing bias is critical for the correct use of an AI system (→ TRANSPARENCY) (see for instance [Ranard et al., 2024](#); [Yang et al., 2024](#)).

## Risks related to insufficient robustness / resilience and generalisability

### **Parent terms:** Non-maleficence – Generating and evaluation evidence on risks

Robustness of AI systems, i.e. their intrinsic or built-in *resilience* to maintain performance also under variable conditions of use, constitutes still a major challenge for AI systems and is one of the main factors affecting trust and hence adoption.

In different AI domains, there are different interpretations of what precisely constitutes robustness. Tocchetti and colleagues (2022) analysed the literature on 'robust AI' and found that robustness is used in conjunction with a variety of other terms, including *noise*, *perturbations*, *distributional shifts* (→ *distributional drift / shift*), *adversarial robustness*, *adversarial attacks*, *generalisation*, → **FAIRNESS**, → *interpretability and explainability*.

- While it is not surprising that robustness can touch on a variety of other concepts, great care should be taken when talking about robustness to describe the precise *notion* one has in mind.

Tocchetti et al., also proposed three taxonomies on robustness: 1) robustness by methods and approaches in different stages of AI development; 2) robustness for specific model architectures, tasks, and systems; 3) robustness assessment methodologies and, particularly, possible trade-offs with other trustworthiness properties (→ inherent incompatibilities and trade-offs).

Apart from the technical aspects listed above, robustness and resilience are closely related to the following concepts:

- **The ethical demand of "prevention of harm"** ([EU HLEG, 2019](#)), here discussed as **non-maleficence** and, in a broader context, → **AI safety**
- **The concept of → reliability**: both robustness/resilience and → reliability ascertain whether an AI system reaches a certain (desired minimum) level of performance under different conditions than those it was developed for. Changed conditions may relate for example to changed input data or data parameters that may result from changed use conditions, → **use environments** or → **use contexts** or adversarial (cyber)attacks or result from drift (→ **concept drift / shift**, → **data drift / shift**, → **distributional drift / shift**).

AI systems should be developed with variability of real-world use conditions and related risks in mind, leading to models that are sufficiently robust or resilient.

Here we focus mainly on

- a) Variations of → **use contexts**, → **use environments** and issues of → **drift / shift** in machine learning.
- b) Cyber attacks can be considered extreme variations of circumstances. AI models should be sufficiently robust and resilient in regard to cyber attacks. This means that any cyber vulnerabilities should be minimized to the extent possible (see [Reina & Griesinger, 2024](#) for an ontology of relevant terms on cybersecurity in the health domain). Importantly, cybersecurity measures must cover all relevant infrastructure in a given → **use environment** (e.g. a hospital), not simply individual devices.

The lower-level concepts below provide more details.

## Robustness / resilience: use context and use environment variations & drift issues

### **Parent terms:** Non-maleficence – Generating and evaluation evidence on risks

#### **Use context variations:**

Variations of → **use context** are obviously a challenge for the performance of an AI system as intended. For instance, 'off-label' use of a specific AI-enabled diagnostic device for a risk group (e.g. rare diseases) for which it was not developed, may carry risks of inadequate → **model performance** or → **accuracy**.

Risks in this regard may be minimized by sufficient → TRANSPARENCY, e.g. by providing

- Clarity on intended → use context.
- Clarity on → intended use and clear → instructions for use
- Outlining conditions of → foreseeable misuse that should be avoided.
- Providing a clear framing of → applicability and limitations.

In addition, approaches to address ‘human factors’ ([Borsci, 2018; Borsci & David, 2020](#)), including through so-called ‘*human factors engineering*’ may reduce risks regarding safety and reliability ([Mishra et al., 2024; Nerincx N & Lindenberg J, 2005](#)).

#### **Use environment variations:**

AI systems that are rolled out will not hit uniform → use environments. Inherently there will be variations of available infrastructure, of workflows, of available skills that are necessary for appropriately using AI systems and in relation to the availability of multi-disciplinary teams that may ideally be required for running and monitoring the AI system.

[Wu et al. \(2021\)](#) show that the → model performance of a pneumothorax AI prediction model can vary by up to 10% from one hospital to another, i.e. with the specific → use environment. Thus, even seemingly small changes in use environment may affect performance and → AI safety (increasing risks for false predictions) and may reduce benefits ([Wu et al., 2021](#)).

AI systems ideally should be designed and deployed in a way to minimise impacts of → use environment variations. This involves various considerations including interoperability, alternative IT infrastructure (see also → risks relating to technical integration and interoperability issues), training material, training requirements, → usability, user interface, clear descriptions and documentation as well as relevant warnings.

#### **Drift / shift**

A challenge for AI systems are the various types of drift or shift that can occur progressively over time or, in some instances, abruptly (→ drift/shift in machine learning). Drifts deteriorate performance and may hence pose significant risks for patients.

While → concept drift / shift and → data drift / shift are difficult to control, a higher degree of resilience to these types of drifts right from the outset would be desirable (see [Kagerbauer et al., 2024](#)). This is an area of ongoing scientific and technical inquiry. At least, potential drifts need to be monitored (→ monitoring drifts / shifts).

Drifts due to changes in → use environments and/ or → use contexts (→ distributional drift / shift; also referred to as ‘deployment bias’) can to some extent be controlled by providing sufficiently transparent information to deployers and users (see → risks related to transparency) and by regular interaction with users (surveys) etc. concerning both → usability and real-world use.

#### Generalisability: shortcut learning

**Parent terms:** Non-maleficence – Generating and evaluation evidence on risks

#### **Shortcut learning poses a considerable risk**

Machine learning approaches using → deep learning (see also → artificial neural networks) may exhibit the phenomenon of → shortcut learning. This means that the model is acquiring, during the training process, ‘shortcuts’ between data features and outputs: instead of utilising for instance relevant features with *predictivity*, it privileges other features that are coincidentally and readily present in the → training data but are irrelevant or less relevant for the purpose (e.g. lack of → conceptual relevance). Shortcut learning is an undesirable consequence of → objective functions and optimisation of models. Shortcut learning will lead

to inaccurate predictions under real-life conditions and lack of → generalisability. It constitutes a considerable risk for → AI safety and, consequentially, → patient safety in case of AI systems used in healthcare.

### **Detecting shortcut learning through lack of generalisability**

To detect shortcut learning prior to deployment of an AI system, rigorous testing of models, e.g. through external → model validation or through → model evaluation should be conducted. This will show issues of lack of → generalisability (i.e. lack of accuracy of the model when faced with real-world data, e.g. data from other patient populations). Shortcut learning may also be detected in the post-deployment space, e.g. through audits and inquiries (→ auditability and auditing).

### **Mitigating the risk of shortcut learning: maximising intelligibility**

The risk of shortcut learning underlines the importance of model → intelligibility: if we understand how and why models produce their outputs (→ output and output data), we will be able to tell whether a model makes the right decisions for the right or for the wrong reasons. Intelligibility entails several elements, most importantly interpretability of the model functioning and, where this cannot be readily achieved (e.g. deep learning-based models) post-hoc explanations of, for instance, key features the model uses of producing outcomes (so-called explainability) (→ interpretability and explainability).

### **Avoiding shortcut learning: choice of AI technique**

The best option to minimise the risk of shortcut learning is to aim for inherently intelligible models ([Weld & Bansal, 2019](#)) when deciding on the → AI technique (i.e. at the outset of the → evidence pathway for AI in health): traditional → machine learning approaches (e.g. random forest, decision trees etc.) do not show shortcut learning and, in addition, typically have a high degree of intrinsic → interpretability, which supports model → intelligibility. Thus, avoiding black-box models for medical applications (see also comments by [Rudin, 2019](#) and [Vokinger et al., 2021](#)) where feasible is a safe strategy to avoid costly issues with models that are non-generalisable or underperforming under real-world situations. A recent systematic analysis of the medical literature shows that clinical research groups prefer complex black-box models while not reporting sufficiently on issues related to intelligibility of the models ([Chassaigne et al., in preparation](#)).

## Cybersecurity: risks for value chain assets and knock-on effects on safety

**Parent terms:** Non-maleficence – Generating and evaluation evidence on risks

### **Cybersecurity as a value-preserving enabler**

According to our definition of the → value chain of AI, we distinguish (see [Reina & Griesinger, 2024b](#)) within the → AI value chain

- enablers / prerequisites and
- assets / values

In this context, the assets / values used and/or created correspond to → machine learning models, to → AI systems and associated services for using AI in health (→ use of AI systems in health and healthcare). An example is telemedicine using → AI systems or → AI-enabled medical device software.

Enablers are a) enabling IT infrastructure, b) enabling technologies, c) cybersecurity and d) data. Among these, we consider cybersecurity ([ENISA, 2016](#)) as a horizontal ‘value-preserving enabler’, protecting all other value chain elements from deterioration / destruction / theft etc. that might result from cyber attacks.

For an ontology of about 70 terms relating to value chain enablers with a particular view to cyber incidents in the health sector, see [Reina & Griesinger, 2024](#). The ontology is based on various sources, drawing to a large extent on the excellent report on the creation of a taxonomy for the European AI ecosystem by the European Institute of Technology ([EIT, 2021](#)).

### **Cyber attacks and knock-on effects on health**

Cyber attacks can provoke the deterioration or destruction of assets of the value chain such as → machine learning models and/ or statistical models, → AI systems (containing machine learning models) or services, for instance those required for the use of digital health solutions (e.g. wearables under supervision of a health setting) as well as data.

Due to the specifics of the healthcare sector, the deterioration and/or destruction of value chain elements caused by cyber incidents has the potential to cause both serious health effects to patients and safety impacts for healthcare staff. Examples of adverse health effects are delayed diagnostic procedures or delayed surgery, prolonged hospitalisation. For a full list of possible adverse health effects caused by cyber incidents see Reina & Griesinger, (2024). Additionally, cyber attacks may also lead to the leakage of → data (including sensitive personal health data) that may cause the ‘poisoning’ of → machine learning models or → data, thus inflicting deterioration and longer-lasting damages.

In this perspective the → CIA principles (i.e. Confidentiality, Integrity and Availability; see [Lundgren & Möller, 2019](#); [ISO, 2018](#); [NIST, 2017](#)) constitute a model widely used in information security that should be guaranteed in order have a secure system. Any organisation that wants to protect its data and information systems, should design and implement cybersecurity plans that include rules and safety measures to reduce potential threats to confidentiality, integrity and availability or to any of their combination.

### **AI and cybersecurity**

While AI systems may be the object of cyber attacks, AI is also used for cybersecurity purposes, e.g. by automatically monitoring, detecting and responding to cyber attacks, by analysing large data sets in view of possible anomalies that may indicate an attack or by scanning network traffic to detect malware or fraudulent practices. In addition, AI algorithms can support the detection of vulnerabilities of a software or within an IT system ([Glover, 2024](#)).

At the same time, AI algorithms and in particular generative AI (→ foundation models; → generative AI), can be used maliciously to create adaptive malware that is difficult to detect with traditional security measures ([Glover, 2024](#)). AI will likely completely transform cyber threats in the near future, by allowing the automatic evolution of attack methods within milliseconds ([Gibson, 2024](#)), requiring AI-powered defence mechanisms that analyse, adapt and deploy countermeasures at the same timescale. Possible preparation strategies include ([Gibson, 2024](#)): i) sufficient human oversight in case of fully automated AI systems (→ automation), including human-AI collaboration protocols allowing control of systems and ii) the establishment of cross-border and cross-sector AI defence networks (e.g. “*Collaborative Automated Course of Action Operations*” ([CACAO](#)); see [OasisOpen, 2024](#)).

## Risks related to value chain elements

### **Parent terms:** Non-maleficence – Generating and evaluation evidence on risks

All elements of the → value chain of AI are a potential source of risks. We discuss here risks related to value chain **assets** or **values**, i.e.

- Data (including services, such as data providers)
- Models and AI systems (including services such as model providers, model libraries or repositories)
- Services required for operating the AI system as intended

N.B. Risk related to other value chain elements are discussed in other sections:

- Risks related to **cybersecurity** (value-preserving enabler) are addressed under → cybersecurity: risks for value chain assets and knock-on effects on safety.
- Risks related to **IT infrastructure** and **enabling technologies** are addressed under → risks relating to technical integration and interoperability issues.

### **Data related risks**

Data are a fundamental of AI systems and consequentially any harm or risks associated with data used to train, evaluate or validate a model will be encapsulated to some extent in that model and AI system.

Data related risk predominantly concern various forms of → bias which may have implications for → AI safety and non-maleficence. See → risks related to bias for more details.

Incorrect or false assumptions, outdated knowledge or concepts may be encapsulated in data, for instance via data labelling (→ labels / data labels) and associated → features, → attributes or → proxies. Thus, issues associated with an incorrect or outdated → conceptual relevance may be propagated via data.

Value chain associated data-related risks may originate from situations where developers purchase or acquire data packages that encapsulate such issues and which are either insufficiently documented or have not never been considered, detected and corrected when compiling the data set.

### **Risks related to models of AI systems**

In regard to the value chain, model-related risks may originate *inter alia* from

- the use of models (e.g. from model libraries), which may have insufficiently documented → algorithmic bias. This includes also situations of → federated learning & split learning and → transfer learning using pretrained (large) models.
- use of (pretrained) models that are outdated, e.g. in regard to underpinning scientific assumptions (→ conceptual relevance) or technical properties that may compromise performance and safety
- inadequacies of federated learning. In particular the local training of federated models may prove problematic, affecting the federated learning optimisation process and → model performance of federated models. In case federated models are subsequently used as a basis for specific models (e.g. through → transfer learning), these issues may propagate along the → value chain of AI ([Singosoglou et al., 2023](#)).
- decisions for models (→ algorithm-to-model transition) that are ill-suited for a given task or unnecessarily complex, resulting in risks related to → interpretability and explainability. The latter may cause effects on i) free and informed decision-making by patients (→ ensuring the means for free and informed consent) or ii) lead to issues concerning the patient-physician relationship (→ upholding a trustful patient-physician relationship). This includes automation bias and complacency (see → avoiding automation bias; → avoiding automation complacency): a physician that cannot comprehend outputs of an AI system may, based on the impressive performance, simply decide to uncritically accept these outputs, without sufficiently controlling each output in regard to the specific case at hand (see also → ensuring human agency and oversight).

### **Risks related to services needed to operate AI systems**

There are manifold risks which may relate

- to a lack of quality and risk management culture by the organisation offering the services, e.g.
  - lack of or insufficient cybersecurity measures

- lack of or unclear attribution of responsibilities (see → accountability structures, attribution of (distributed) responsibilities)
  - lack of or unclear communication lines (in particular for collecting and responding to post-deployment problems or user issues)
  - insufficient monitoring of services-related problems for contributing for a full understanding of possible → incidents and → adverse events.
- Lack of sufficient training and skills of staff executing services in relation to the AI system
- Insufficient technical integration between the AI system and the infrastructure required for the services (→ risks relating to technical integration and interoperability issues)

## Risks relating to technical integration and interoperability issues

### **Parent terms:** Non-maleficence – Generating and evaluation evidence on risks

The implementation of AI systems in health presents several risks across various aspects of integration and interoperability. These risks run across the four application domains of AI in health ([WHO 2021](#)): healthcare, medical research, public health surveillance, and health system administration.

Integrating an AI system into novel and, in particular, existing → use environments poses significant technical challenges that may impact on the functioning of AI systems as intended.

Technical challenges relate to → value chain aspects, including communication protocols, data format issues, IT infrastructure and enabling technologies. For an ontology of terms of value chain enablers including cybersecurity see [Reina & Griesinger \(2024b\)](#). For further reading on practical challenges regarding the deployment of AI systems, see [Baier et al., 2019](#) and [Paleyes, 2022](#) and [CISCO website, 2024](#).

Technical issues include:

- 1) **Lack of standardization of communication protocols.** For instance, the lack of robust *Application Programming Interfaces* (APIs) can result in limited functionality of various digital applications, including AI systems due to restricted data access and potential cybersecurity vulnerabilities if APIs are not properly secured.
- 2) Lack of **standardization in data formats** between different AI systems and (legacy) medical devices, AI systems and existing data sources (e.g. for AI research tools) including electronic health record (EHR) systems (in healthcare settings). Public health surveillance typically exploits various data streams (e.g. hospital admissions, social media, and environmental data) which may be difficult to integrate into one system for subsequent analysis, pattern and signal detection.  
**These issues may cause the following issues, which may, alone or in combination, cause adverse events in patients:**
  - incomplete or inconsistent data access (→ input data, → output and output data)
  - data misinterpretation or use of fragmented data (e.g. patient information in healthcare) with impacts on → model performance
  - impaired → model performance and/or → accuracy (e.g. in AI-driven diagnoses and treatment recommendations)
  - inefficient data exchange requiring additional data transformation steps, increasing probability of error
  - increased complexity in maintenance and updates, increasing probability of error
  - privacy (→ PRIVACY PROTECTION) and → data quality issues
- 3) **Inadequately performing and configured (existing) IT Infrastructure.** AI systems pose significant demands on IT infrastructure including *processing capabilities* for complex AI algorithms, *storage capacity* to handle the large amounts of data required for AI training and operation,

*bandwidth requirements and latency issues* (delays in electronic communication). In particular, existing IT infrastructures at a given setting (e.g. research environment, hospital) may not be adequate for integrating a demanding AI system. Further, infrastructure problems include interaction with existing (legacy) medical devices which may be required for operation of the AI system for healthcare. **IT infrastructure issues may lead to:**

- degraded → model performance of AI systems
- inability to process real-time data for time-sensitive applications. For instance, delays in real-time AI applications such remote patient monitoring
- limited scalability of AI solutions with potential impacts on → health equality & health equity
- increased costs for data management and storage solutions
- inconsistent performance of AI systems across different healthcare facilities (→ health equality & health equity)

**4) Enabling Technologies:** The implementation of AI systems in healthcare relies on several enabling technologies that may pose risks if not properly used and/or configured:

- improper implementation of containerization technologies
- challenges in orchestrating AI workloads, impacting workflows (→ clinical pathway, → clinical practice protocol)
- over-engineering of orchestration setups, e.g. due to complexity of AI systems
- inefficient container orchestration and subsequent suboptimal resource allocation
- lack of proper monitoring and scaling mechanisms
- inconsistencies between development and production container environments
- improperly configured network policies in orchestration
- insufficient isolation between virtual machines
- performance overhead introduced by virtualization layers
- usage of inconsistent, outdated or unpatched versions of → machine learning (ML) libraries. This may impact → model development, update, adaptation and may hence degrade → model performance
- usage of generic ML libraries may not be optimized for healthcare-specific tasks

**These issues may result in:**

- inconsistent AI → model performance
- service disruptions during peak loads
- impacts to the real-time capabilities of healthcare AI applications
- unnecessary complexity and potential points of failure, impacting clinical workflows (→ clinical pathway, → clinical practice protocol)
- exposing sensitive healthcare data or provide attack vectors for malicious actors (→ PRIVACY PROTECTION).
- reducing model → accuracy or → efficacy, effectiveness and efficiency

## Risks related to integration into workflows

**Parent terms:** Non-maleficence – Generating and evaluation evidence on risks

When integrating AI systems into workflows, e.g. specific clinical workflows (→ clinical pathway; → clinical practice protocol; → clinical practice guideline), care must be taken to sufficiently consider the benefit-ratio of an AI device and its potential impacts on patients (e.g. for diagnosis, clinical decision making etc.). An AI system, once integrated into a workflow at a healthcare setting and even more so in a health system, has 'multiplier' effects and could quickly affect a high number of patients.

Consideration also must be given to how residual or unavoidable → bias may affect patients or patient groups, to → intrinsic incompatibilities and ‘trade-offs’ as well as → calibration of the AI system (where applicable): these aspects may greatly influence the likelihood of undesired outcomes, if not sufficiently considered.

For risks in relation to issues such as the patient-physician relationship, automation complacency or automation bias, see → DIGNITY, FREEDOM AND AUTONOMY.

### Risks related to pre-deployment evidence gaps

#### **Parent terms:** Non-maleficence – Generating and evaluation evidence on risks

Risks related to evidence gaps in the pre-deployment space are a serious problem. → Biases, errors, → usability issues may go undetected, with potential consequences for patient / → AI safety.

Wu et al., (2021) found that a considerable proportion of AI-enabled medical devices had been evaluated solely on retrospective data, collected before the devices were developed. Retrospective evaluation alone is much more prone to → overfitting the → machine learning model (Zou & Schiebinger, 2021) and should hence not constitute the sole source of evidence before an AI system is deployed for routine healthcare use.

Issues include:

- *For non-healthcare uses:* lack of adequate verification and validation activities, e.g. → AI system validation or → usability validation.
- *For healthcare uses:* Lack of sufficient → clinical evidence (including on risks) due to inadequate or insufficient evidence in prospective studies involving representative patient populations (see → model evaluation, → clinical investigation, → clinical validation, → clinical evaluation).

### Risks related to insufficient post-deployment monitoring / post-market surveillance

#### **Parent terms:** Non-maleficence – Generating and evaluation evidence on risks

Insufficient or inadequate efforts of → post-deployment monitoring or → post-market surveillance carries significant risks in regard to detecting in a timely manner any issues or problems that may arise when a system is used under real-world conditions: these will unavoidably differ from situations of research settings and controlled conditions used when designing, developing, evaluation and validating a system.

- Problems may manifest themselves as → incidents or → adverse events (→ monitoring incidents / adverse events).
- More fundamentally, issues of performance, effectiveness, efficiency and bias should be proactively monitored (→ monitoring performance, effectiveness, efficiency and bias)
- Drifts and shifts may affect AI system / model performance (→ monitoring drifts / shifts).

### Monitoring incidents / adverse events

#### **Parent terms:** Non-maleficence – Generating and evaluation evidence on risks - risks related to insufficient post-deployment monitoring / post-market surveillance

Significant problems may manifest themselves as → incidents or → adverse events. These need to be monitored continuously in the post-deployment space to enable the correction of problems and relevant necessary communication (→ correcting problems and failures, including necessary communication) and to provide transparent information failures and, where applicable their root causes (→ failure transparency). Incident / adverse event monitoring are part of → post-market surveillance.

## Monitoring performance, effectiveness, efficiency and bias

**Parent terms:** Non-maleficence – Generating and evaluation evidence on risks - risks related to insufficient post-deployment monitoring / post-market surveillance

Reliance on monitoring → incidents and → adverse events is not sufficient: it will pick up problems once it is already too late. Instead, issues of performance (see also → model performance), effectiveness, efficiency (see → efficacy, effectiveness and efficiency; see → clinical effectiveness) and → bias should be proactively monitored, e.g. by maintaining close dialogue with users, clinicians and patients.

## Monitoring drifts / shifts

**Parent terms:** Non-maleficence – Generating and evaluation evidence on risks - risks related to insufficient post-deployment monitoring / post-market surveillance

Various forms of drifts should be monitored, e.g. through following the relevant literature, dialogue and surveys with users and data scientists (→ drift / shift in machine learning). There are various notions of drift or shift: → concept drift / shift, → data drift / shift, → distribution drift / shift.

Drift / shift issues may also be detected through performance monitoring and incident/adverse event monitoring (→ monitoring performance, effectiveness, efficiency and bias).

## Monitoring usability and usability-related errors

**Parent terms:** Non-maleficence – Generating and evaluation evidence on risks - risks related to insufficient post-deployment monitoring / post-market surveillance

Problems with → usability, i.e. the ease and facility of use by intended users, can pose significant risks for patients. Organisations developing and/or deploying AI solutions should be in an active dialogue with users to proactively monitor and any potential usability issues so that these can be corrected.

This includes also adjustment of training requirements and training programmes where needed (→ information on training requirements). Consideration should be given to variations of → use environments, e.g. when using AI systems in resource-constrained settings ([Konduri et al., 2018](#)) or in situations of home care / domiciliary use ([Black, 2023](#)).

## Generating evidence on wider impacts

**Parent term:** Non-maleficence

NON-MALEFICENCE is not only concerned with harms and risks for persons, but also with potential (harmful) impacts of AI systems on groups, communities and society. Such impacts can be assessed *ex ante* and *ex post* through dedicated ‘impact assessments’ that may be designed on the basis of standardized approach with adaptations as required by the specific AI system, its application in health etc. See → AI impact assessment (AI-IA), → fundamental rights and algorithm impact assessments.

## Impact assessments

**Parent terms:** Non-maleficence – Generating evidence on wider impacts

Relevant concepts of impact assessments are:

- → Ethical evaluation of AI
- → AI impact assessment (AI-IA)
- → Fundamental rights and algorithm impact assessments
- Audits of organisations involved in AI development, deployment and use and of AI systems (→ auditability and auditing).

## A.3 Dignity, freedom and autonomy

### Concept description

#### Dignity, freedom and autonomy

Briefly, **human dignity refers to the principle that every human being has an inviolable absolute worth** (Latin “*dignitas*” = worth), independent of any other individual features or properties, whose value may vary between societies or groups and/or be subject to change over time (e.g. wealth, gender, education, physical aspects, health status etc.). Since disease renders patients particularly vulnerable, it is essential that healthcare upholds the dignity of patients. Dignity, freedom and autonomy are closely related:

The modern concept of human dignity was influenced to a high degree by the works of Immanuel Kant (*Groundwork to the Metaphysics of Morals*, GMM (1785)): human dignity is a status which places the life of human beings above all price (i.e. having no price). Kant uses the term *Sittlichkeit* or *Moralität (morality)* of mankind and individuals to identify the one and only fundamental that lends **dignity** to everybody: dignity is rooted in morality (GMM§435). Morality is a condition for an individual to be not only means but *purpose* (GMM §433). Purpose in itself (*Zweck an sich selbst*) is posited as constraint on **freedom** (GMM §431) – freedom to will or act in agreement with the maxim that acts could be general law (the often cited ‘categorical imperative’, GMM§432). This congruence of *individual* yet *general* ‘law making’ is referred to as **autonomy** (*Autonomie*). Kant contrast it with heteronomy which binds individuals merely by duty to laws, without ever being able to deduce the origin of duty in this context (GMM§431).

Thus, in the **philosophy of the European enlightenment**, human **dignity is closely intertwined with fundamental human rights, such as → autonomy and freedom** (from causality, i.e. freedom of choice) as well as ‘personhood’. According to the EU charter on fundamental rights (ECHR, 2000), dignity is both a fundamental right and the real basis for other fundamental human rights (Article 1 of ECHR, 2000). Human dignity is a foundational concept also in the European Convention on human rights (Council of Europe, 1950) and the United Nations international covenant on civil and political rights (United Nations, 1966).

A key international document of dignity in the (bio)medical context is the **Oviedo Convention on human rights and biomedicine** (ETS No. 164) by the Council of Europe (Council of Europe, 1997a,b; for an analysis in view of AI, see also Mittelstadt, 2021). The **foundation principle** of the Oviedo convention is **human dignity**, from which the convention refers to other values and rights, e.g. **prohibition of discrimination** or the **right to privacy**. Given both the philosophical background to these European Conventions and these Conventions, dignity – in this ontology – is linked to →PRIVACY PROTECTION as well as → FAIRNESS).

#### Dignity, freedom and autonomy in the context of biomedicine

Dignity, freedom and autonomy are of fundamental importance in **biomedicine** from various perspectives: personal dignity includes **free and informed consent**, i.e. the freedom of the patient to choose medical treatments autonomously and to give, withhold or withdraw consent to a treatment; it also includes the **freedom to know about one's medical condition as well as the freedom or right not to know** (known under the acronym “RNTK” in medicine). Dignity is also linked to the respect for **private life in regard to medical information** (→ medical privacy / health privacy), including the freedom to give or withhold consent in regard to the collection, processing, storage and use of information about one's health (e.g. for electronic health record and/or biomedical research). Dignity is a key foundation of the **patient-physician relationship** and encompasses the **dignity of both, patient and physician** (Emanuel & Emanuel, 1992; Mittelstadt, 2021).

What concerns → AI principles and AI ethics guidelines, dignity, often linked to freedom and autonomy, is typically not further defined (Jobin et al. 2019). However, some international guidance documents on AI discuss dignity in relation to AI in general (e.g. EC HLEG, 2019) or in relation to health (WHO, 2021 & 2024).

Notably, the principle of dignity is reflected in several recitals of the EU's AI Act ([EU, 2024a](#)), e.g. 27, 28, 31, 48, 58, 80.

### Explanatory note

With the advent of AI in medicine and healthcare, there are novel challenges for the dignity of patients, but also healthcare professionals. For instance, black-box AI tools whose outcomes are not sufficiently intelligible (→ [intelligibility](#), → [interpretability and explainability](#)) or are not understandable for a non-medical expert due to a lack of → [understandable explanations](#), would severely affect a patient's right for informed consent. Yet, many clinical research groups developing AI solutions choose inherently inscrutable or non-intelligible AI ([Chassaigne et al., 2025](#)). Lack of → [intelligibility](#) also erodes the physician's professional dignity and thus may damage the trust that is essential for a functioning patient-physician relationship ([Mittelstadt, 2021](#)), including models of patient-physician interaction that place emphasis on patients' values and interests ([Mittelstadt, 2021; Emanuel & Emanuel, 1992](#)). While the EU's AI Act ([EU, 2024](#)) does not ban the use of black-box AI, it addresses the issue through means of documentation (see also → [transparency of AI systems](#)) and → [human oversight](#) ([Panigutti et al. 2023](#)).

Other challenges include the increasing digitalisation of health information, resulting in increased risks concerning the collection of and potential uncontrolled dissemination of personal health information. The EU's GDPR addresses these issues by emphasising that individuals' consent is required for the collection and processing of personal information, including health information.

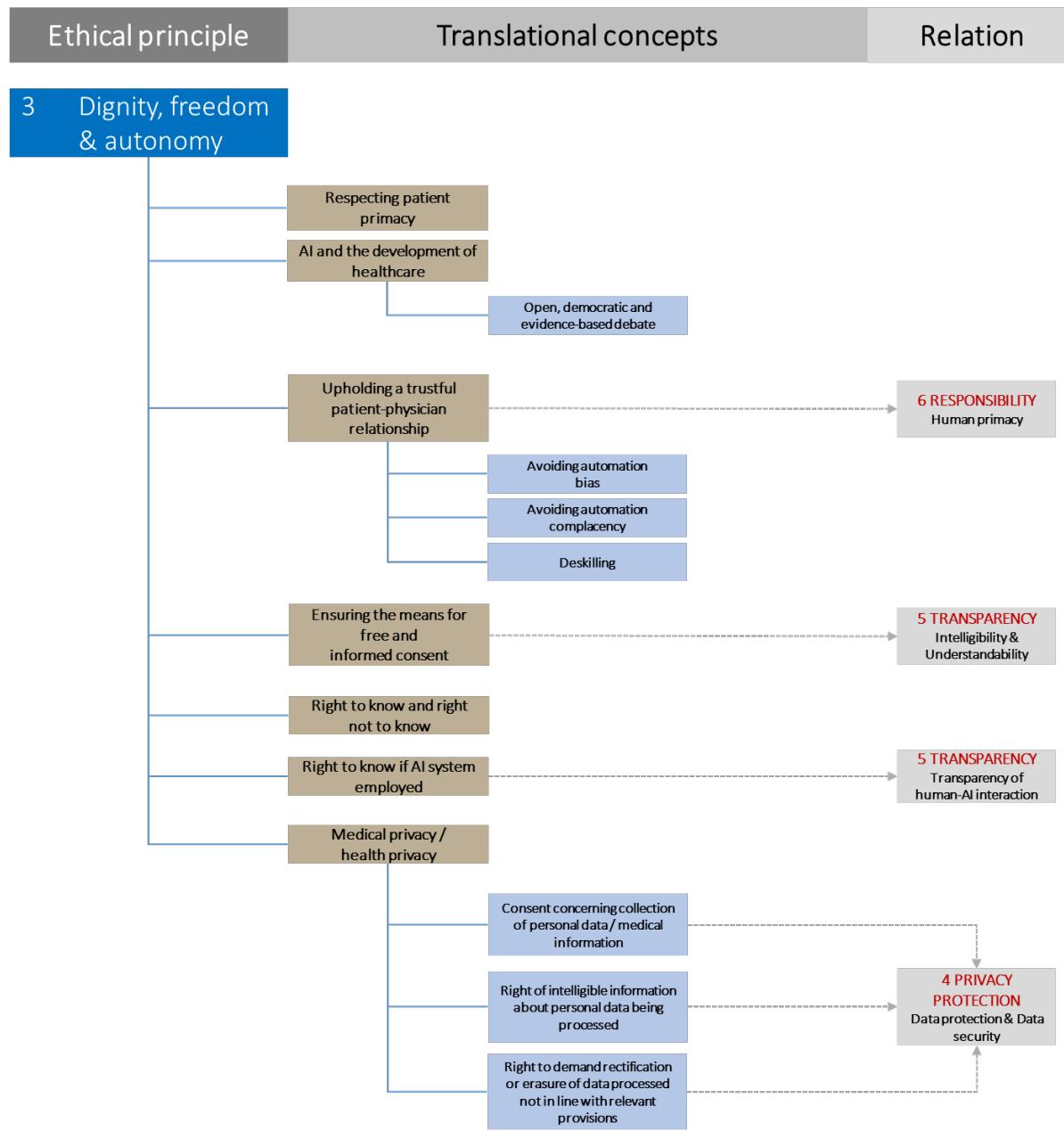
The WHO's guidance on ethics and governance of artificial intelligence for health discusses dignity in several contexts, including health data collection and use or predictive analytics in healthcare (e.g. the right to know and not to know as established in the Oviedo convention).

### Term relationship

Related terms:

- FAIRNESS
- Human agency
- Autonomy
- Patient primacy
- Human primacy
- Intelligibility
- Interpretability and explainability

Ontological organisation of the ethical principle of “dignity” and its translational concepts. Relations to other translational concepts are indicated in grey / stippled arrows.



## Dignity: translational concepts

### Respecting patient primacy

**Parent term:** Dignity, freedom & autonomy

The Oviedo Convention (Article 2) (→ DIGNITY, FREEDOM AND AUTONOMY) outlines the “**Primacy of the human being**” as follows: “*The interests and welfare of the human being shall prevail over the sole interest of society or science.*” Protecting the primacy of patient interests (→ patient primacy) over societal, scientific, economic, socioeconomic or other interests should be a fundamental priority in healthcare settings. This relates in particular to situations where disregarding this primacy might lead to harm of a patient, violating also the bioethical principle of → NON-MALEFICENCE. AI systems in healthcare must not erode this primacy.

Possible issues include:

- The collection and/or use, without appropriate patient consent and/or without sufficient safeguard to protect privacy (e.g. through pseudonymisation), of personal digital health data for scientific purposes or to train AI systems, for example under misguided premise of the “greater good” or “common interest” (see also [McNair et al., 2019](#); [EU HLEG, 2019](#); [WHO 2021](#), [OECD 2024](#)).
  - **Who may be concerned?** Companies and healthcare settings (e.g. hospitals) developing AI, health researchers, medical associations
- Ways of implementing AI systems in healthcare settings in a manner that puts population needs above individual patient needs (e.g. automated decision-making tools, triage systems, public health tools).
  - **Who may be concerned?** Hospitals and other healthcare settings developing AI, health researchers, medical associations, healthcare organisations and healthcare system administrators
- Development and/or use of AI systems (e.g. for diagnosis) with insufficient regard to patient-centric approaches, e.g. by giving preference to universal applicability and/or economic considerations of deploying “one-size-fits-all solutions” over approaches tailored to specific patient needs or to specifically vulnerable groups (e.g. for rare disease diagnostics. For a review on the potential of AI/ML for improving diagnostics for rare diseases, see: [Abdallah et al., 2023](#)). See also chapter 1.5 on diversity, non-discrimination and fairness” and the subpoint on “universal design” in EU high-level expert group guidelines on trustworthy AI of 2019 ([EU HLEG, 2019](#)).
  - **Who may be concerned?** AI developers, hospitals and other healthcare settings.

### AI and the development of healthcare

**Parent term:** Dignity, freedom & autonomy

AI has a tremendous potential to improve healthcare, but there are – as with any technology – also considerable risks. These are influenced mainly by two factors:

- The **AI technology itself and how AI is developed**, respecting ethical principles such as inclusiveness, non-discrimination, equity and/or equality (→ FAIRNESS). The rapid technological developments may create further use scenarios, harms and risks.

- The way AI is implemented and used in health and medicine. While it can be used to speed up diagnoses, reduce heuristic decision-making, enhance accuracy, relieve burden of healthcare professionals and so on, it can also be (mis)used. Areas of concern include:
  - **Coercive practices / “nudging”:** For instance, coercive practices concerning behavioural changes (keyword “big nudging”) involving for instance → **conversational agents** in healthcare ([Laranjo et al., 2018](#)) raise numerous ethical questions in regard to dignity, autonomy, freedom of patients. These ethical risks would be obviously exacerbated in cases where (wearable) medical devices are connected to the internet and covert use of these data (e.g. for “training” of physicians, for training AI systems or, particularly problematic, to inform insurances). We refer to the excellent review by Arnold for a detailed discussion of “nudging” by AI/robotic agents ([Arnold et al., 2021](#)) and other pertinent ethical considerations for use of AI systems in healthcare.
  - **Economisation of healthcare:** Similarly, AI may lead to a “economisation” of healthcare, neglecting sufficient consideration of individual circumstances, patient history, preferences etc. and giving too much regard to cost-benefit calculations on a population level, thus violating the patient primacy over societal and other interest (→ **Respecting patient primacy**).
  - **Use of → conversational agents (‘chatbots’):** use of chatbots (e.g. for patient workflows or treatment of mental diseases and conditions) needs to be closely monitored in regard to experiences of patients and users, including potential effects on mental health. In particular affecting computing may not only provide benefits, but also pose risks. For instance, the use of anthropomorphized computer agents has shown to lead to adverse behavioural effects ([Brahnam, 2006](#); [Darling, 2016](#)).

## Open, democratic and evidence-based debate

**Parent terms:** Dignity, freedom & autonomy – AI and the development of healthcare

The use of AI in healthcare, as an application field with high ethical, technical, privacy-related and clinical risks, requires a **continuous and open and democratic debate** that is **evidence-based**, i.e. informed by trustworthy scientific, clinical information as well as information from users and stakeholders based on their experiences. N.B. this is not a point specific to AI: other new biomedical technologies should be and are also subject to continuous debate, e.g. gene-editing techniques.

Such debate should also take into consideration legitimate concerns regarding a complex nexus of ethical questions around dignity, touching on topics such as **societal control**, the danger of a **“dehumanisation” or “mechanisation” of medicine**, its **“economisation”**, which we understand as the risk that medicine morphs from (a perhaps idealised image of) a relationship of care for the vulnerable into a business with insufficient regard to the primary goal: to prevent disease, to cure, treat and heal (→ **patient primacy**).

For further detailed considerations, we refer to the WHO guidance on ethics and governance of artificial intelligence for health ([WHO, 2021](#)) and in particular to the section *“emerging trends in the use of AI in clinical care”* which distinguishes four major trends imposed by AI: a) the evolving role of patients in clinical care, b) a shift from clinical to home-based care, c) use of AI for clinical care in contexts outside the formal health system and d) the use of AI for resource allocation and prioritisation, including in critical situations (e.g. pandemics).

- **Who may be concerned?**

**All stakeholders along the AI development and use pathway**, in particular **health system organisations, health system managers, policy makers, clinicians, medical associations and patient associations**.

*An important element for effective open democratic debates is not only sufficient information but the understandability of the issue at stake to those that do not have a medical, scientific, bioethical background.*

## Upholding a trustful patient-physician relationship

**Parent term:** Dignity, freedom & autonomy

A trustful relationship between patient and physician is a fundamental pillar of the traditional view of medicine as a practice of healing. Medicine's "*most cherished and defining values including care for the individual and meaningful physician-patient interactions*" may suffer from "McDonaldization", i.e. the fixation on key principles *efficiency, calculability, predictability, and control* ([Dorsey & Ritzer, 2016](#)).

While the reality in modern healthcare settings and health systems, e.g. due to economic pressures, lack of staff etc., may not correspond to this idealised relationship, it remains a strong *desideratum*. It roots in the specific vulnerability of patients because of a compromised health (→ **DIGNITY, FREEDOM AND AUTONOMY**).

We refer to the excellent report by [Mittelstadt \(2021\)](#) on the impact of AI on the doctor-patient relationship commissioned by the Council of Europe's steering committee for human rights in the fields of biomedicine and health (CDBIO) as well as the excellent analysis of Arnold, addressing a variety of ethical topics of AI in medicine including the doctor-physician relationship ([Arnold, 2021](#)).

A recent poll by the Pew Research Center found that, in the US, most patients fear that the use of AI systems in healthcare would negatively impact the patient-physician relationship ([Pew Research Centre, 2023](#)).

We list some key issues and considerations below:

- **Denigration of the physician's skills and erosion of patient trust:** By reducing human agency and the apparent value of human agency (e.g. based on performance or efficiency advantages of an AI system), the skills, experience and training of the physician may be denigrated, which will erode the trust of the patient in the doctor's capacities. This may even hold in cases where AI systems are used under conditions of human oversight (→ [augmentation](#)) and where AI systems are merely used to enhance the efficiency and / or accuracy of the physicians. It should be noted that distrust in medicine and physicians is not a new phenomenon ([Arnold, 2021; section "Trust"](#)), but the use of algorithms, the personification of technology ("iDoctor"; see [Karches, 2018](#)) and a depersonalisation of care (driven for instance by economic benefits of the employment of AI systems) may accelerate this further.
- **Lack of attunement to individual patient needs:** consideration of individual patient needs and patient-specific circumstances (e.g. based on a patient's health history, experiences with treatments, exam results, choices and preferences) requires attunement of the physician to these factors ([Karches, 2018](#)). Strong reliance on AI systems and resulting data representations ([Mittelstadt, 2021](#)) may interfere with this human dimension of healthcare, leading to frustration and deterioration of the perceived quality of the patient-physician relationship, in itself perhaps an important factor for the process of "healing" (→ [avoiding automation bias](#); → [avoiding automation complacency](#)). Having said this, AI systems may help reducing → [heuristics](#) in clinical decision-making by providing unprecedented speed and accuracy of the integration of complex information from various sources (genetic tests, biomarkers, radiology etc.). The advent of AI systems able to process multi-modal data (→ [data modality](#)) may further accelerate this trend ([WHO, 2024](#)). However, this should not come at the cost of personalised care centred at the individual.
- **Replacement of physical patient-physician interaction with virtual systems:** Deployment of AI systems will likely not happen uniformly across healthcare settings, regions, countries and global regions. It is conceivable that remote (telemedicine) AI systems (e.g. chatbots, → [conversational agents](#)) may be used to tackle lack of coverage of healthcare access ([Mittelstadt, 2021](#)) in

various countries. While such systems clearly tackle gaps of healthcare coverage, they will not replace a personal interaction between patient and physician (see also: → FAIRNESS - health equality & health equity) and there is a risk that AI systems may cement existing or worsening inequalities, including non-availability of a personal patient-physician relationship. In addition, chatbots or virtual assistants are increasingly proposed for various treatments (e.g. rehabilitation, neuropsychiatric disorders). The impacts on patient dignity, autonomy, health and safety (→ AI safety) and -privacy (→ PRIVACY PROTECTION) are not fully clear yet and are difficult to regulate (Ebers, 2024a).

- **Who may be concerned?**

*Actors and organisations that decide how AI systems are used in healthcare, e.g. health system organisations, health system managers, policy makers, clinicians, medical associations and patient associations.*

## Avoiding automation bias

**Parent terms:** Dignity, freedom & autonomy – Upholding a trustful patient-physician relationship

Automation bias happens when actors give *a priori* preference to outputs (e.g. recommendations) of AI systems or other automated results. Automation bias can be seen as a kind of → heuristics (“anchoring heuristics”). Automation bias is clearly not only an issue of dignity, but may also carry substantial clinical safety risks (e.g. Khera et al. 2023; Challen et al., 2019) (→ AI safety; → NON-MALEFICENCE).

- **Avoiding “programmed” automation bias in clinical guidelines:** When implementing AI systems and, in particular, when integrating AI solutions in workflows, → clinical practice guidelines, → clinical practice protocols or → clinical pathways, care should be taken to avoid preferring their outputs *a priori* as superior to other information sources or processes of reasoning. This is particularly relevant in cases where insufficient information on → real-world benefits is available (“giving AI the benefit of the doubt”, based on promising or impressive specifications or an uncritical general trust in technology) (Enid et al., 2009).
- **Ethical conundrums and automation bias:** Automation bias may be particularly likely in cases where no clear benchmark or cut-point of normality versus abnormality is available, e.g. in situations of ‘ethical conundrums’ (Arnold, 2021).
- **Automation bias and propagation of existing biases in medicine:** Without looking out for and reducing automation bias there is also a great risk to propagate existing “traditional” medical biases in AI systems, including those affecting specific patient groups (Straw, 2020), which makes such biases also an issue of → FAIRNESS.

There are various entry points of such biases in AI models, from data selection and data processing when compiling → development data (Liu et al., 2022), → feature selection, → labels / data labels, biomedical assumptions, hypotheses and theories (→ conceptual relevance; → valid clinical association / scientific validity) such as the predictive relevance of a specific biomarker to algorithmic fine-tuning (→ algorithm-to-model-transition).

- **Automation bias as a chance to question fundamental biases in medicine:** Inversely, the heightened attention concerning biases that comes with debates about AI in health is an opportunity to “interrogate” medical practice regarding engrained medical biases and question the validity and evidence base of long-held assumptions / “truths”. Eliminating biases in healthcare is not only an academic exercise but involves updating medical training (Vela et al., 2022)

- **Who may be concerned?**

*Medical associations, physicians, healthcare organisations, patient organisations, in particular those representing patients with rare diseases.*

## Avoiding automation complacency

**Parent terms:** Dignity, freedom & autonomy – Upholding a trustful patient-physician relationship

**Automation complacency** can be considered a specific form of automation bias ([Parasuraman & Manez, 2010](#)). Automation complacency happens when human actors assume that AI systems (or other automated systems) operate better, e.g. have higher diagnostic → **accuracy**, and make less errors ([Arnold, 2021](#)). This complacency could affect the level of → **agency** of health-care professionals and their engagement in patient care. It may also result in reduced scrutiny of technologies or and reduced human oversight (see → [ensuring human agency and oversight](#)) of AI systems.

Thus, **automation complacency reflects an *a priori* assumption that the automated system is in all cases superior to the human and reduces necessary safeguards**. This may lead to reduced critical inquiry concerning the quality of outputs, including detection of → **biases** or reduced expectations concerning → **interpretability and explainability**, resulting in scientific explanations being replaced by an assumed “**automation superiority**”).

- **Safety implications:** Automation complacency has implications for → **AI safety** (see → **NON-MALEFICENCE**) ([Challen, 2029](#); [Khera et al., 2023](#)). It may lead to insufficient consideration of other information sources or other avenues of treatments (in case of AI systems supporting clinical decision-making), which could lead to patient harm ([Challen et al., 2019](#)).
  - **Fairness and equality implications:** It may have implications for fairness and healthcare equality and disparity (→ **FAIRNESS – health equality & health equity**) ([Straw, 2020](#); [Vela, 2022](#)).
  - **Implications for patient preferences and values:** Finally, complacency and affect the respect for patient preferences and values through perceived “**automation superiority**”.
- **Who may be concerned?**  
*Medical associations, physicians, healthcare organisations.*

## Deskilling

**Parent terms:** Dignity, freedom & autonomy – Upholding a trustful patient-physician relationship

AI systems in medicine may be implemented to enhance the efficiency and performance of skilled physicians and the quality of clinical workflows (e.g. when used as a tool for → **augmentation** or for → **reducing heuristic decision making** during clinical decision-making processes).

However, if implemented in an uncritical manner, AI systems may, on the mid to long term contribute to a loss of skills due to overreliance on AI ([Duran, 2021](#); [Arnold, 2021](#); [Aquino et al., 2022](#); [Sambasivan & Veeraraghavan, 2022](#); [LeLagadec et al., 2024](#)), e.g. if radiological experts delegate tasks to AI systems without keeping up a routine practice of interpreting and analysing medical images so as to re-train their pattern recognition capacities for diagnostic decision-making.

Finally, AI may not only degrade skills but also undermine the “epistemic authority” ([Grote & Berens, 2019](#)) and the self-esteem of healthcare professionals by degrading their ability to influence or at least to fully understand their work, in particular in case of black-box AI with insufficient or unsatisfactory levels of intelligibility (→ **TRANSPARENCY – intelligibility and explainability**). This may erode the care relationship between patient and physician.

- **Who may be concerned?**

*Actors and organisations that decide how AI systems are used in healthcare, e.g. health system organisations, health system managers, policy makers, clinicians, medical associations and patient associations.*

## Ensuring the means for free and informed consent

**Parent term:** Dignity, freedom & autonomy

### ***Free and informed consent: basic meaning and key international conventions***

Free and informed consent means that patients must agree to interventions before these are carried. Patient agreement must be based on sufficient and sufficiently *clear* information ('informed') and must not be subject to undue pressure ('free'), e.g. from healthcare professionals, organisations or other persons.

Free and informed consent implies that, in the context of AI-enabled healthcare products, patients are not subject to fully automated decisions on treatments by an AI system without the possibility to provide his/her/their views.

This requires that patients know (1) to which extent AI systems are being employed by their physician and healthcare provider for decision making and that (2) they are aware of the fact that AI systems are employed (→ Right to know if AI system employed). Healthcare professionals thus must consider how to communicate with patients about the contributions of AI tools, e.g. diagnostic or predictive models that augment (→ augmentation) the clinical decision-making process ([Martinez-Martin, 2018](#); [Schiff & Borenstein, 2019](#)).

In Europe, these two aspects are reflected in two fundamental conventions: the **Oviedo Convention** and the **Convention for the protection of individuals with regard to the processing of personal data**.

#### a) [Oviedo Convention on human rights and biomedicine](#)

Article 5 of the **Oviedo Convention** ([Council of Europe, 1997a](#); see also entry → DIGNITY, FREEDOM AND AUTONOMY) affirms the right to free and informed consent prior to medical interventions or medical research: "*An intervention in the health field may only be carried out after the person concerned has given free and informed consent to it. This person shall beforehand be given appropriate information as to the purpose and nature of the intervention as well as on its consequences and risks. The person concerned may freely withdraw consent at any time*".

Free and informed consent requires that patients are not unduly pressured into specific (diagnostic or treatment) decisions by their physicians. Patients may value their quality of life higher than diagnostic clarity or the side effects of a given intervention. In any case patients must have sufficiently understandable facts at hand to make truly informed decisions. The **Explanatory Report to the Oviedo Convention** ([Council of Europe, 1997b](#)) states that the requirement for consent "makes clear patients' autonomy in their relationship with health care professionals and restrains the paternalist approaches which might ignore the wish of the patient".

Further, Article 36 of this report states: "*The patient must be put in a position, through the use of terms he or she can understand, to weigh up the necessity or usefulness of the aim and methods of the intervention against its risks and the discomfort or pain it will cause*".

Thus, even under modernised versions of traditional "paternalistic" models of patient-physician relationships ([Emanuel & Emanuel, 1992](#)), treatment considerations need to be justifiable and hence explainable, i.e. patients require clear, sufficiently detailed information of reliable quality as a prerequisite to exercise their right for free and informed consent.

Having said this, it is far from clear what concerns a sufficiently understandable explanation in medicine and whether at all truly "mechanistic" explanations can be achieved in medicine ([London, 2019](#)). As summarised by [Herzog \(2022\)](#): "*In medicine fully mechanistic explanations may not be available, while practitioners and patients alike might be satisfied with invoking correlative evidence as the foundation for responsible shared decision making. In addition, the question of "how does it work?" may also not relate to concrete evidence leading to specific decisions (or rather suggestions), but rather to the explanation of fundamental procedural mechanisms employed and conditions relevant to arrive at them. This may entail providing information about the data bases and its provenance as well as interpretive processes and algorithmic processing involved to reach an output.*"

Thus, the exact level of satisfactory explanations that are deemed sufficiently causal and clear (→ explanations of AI systems and their outcomes; → understandable explanations) will depend on → use context

and other factors, such as the level of scientific and medical knowledge about specific disease aetiology / causation and the individual and specific circumstances of the patient, including the stage of disease progression.

For additional points, see → TRANSPARENCY – intelligibility.

a) Convention for the protection of individuals with regard to the processing of personal data

Article 9 (Rights of the data subject) of the **Convention for the protection of individuals with regard to the processing of personal data** (Council of Europe, 1981, modernised in 2018) stipulates that „*Every individual shall have a right ... not to be subject to a decision significantly affecting him or her based solely on an automated processing of data without having his or her views taken into consideration*“.

### **Lack of intelligibility: a challenge for free and informed consent by patients**

Lack of → intelligibility obviously challenges free and informed consent: Black-box AI represents AI systems that are not intelligible), i.e. whose functioning is not intrinsically interpretable, and it is unclear why and how they produce given outputs (→ interpretability and explainability). Why this may not be an issue in specific technical applications such as aircraft collision avoidance ([Burkart & Huber, 2023](#)), it is an unacceptable situation in healthcare since AI-augmented medical decision making (→ augmentation) could not be fully explained and hence justified by healthcare professionals vis-à-vis patients. Thus, without adequate measures allowing to obtain intelligible outcomes and communicate these understandably to patients (→ understandable explanations), such → AI techniques conflict with the need for informed consent by patients to treatments and other medical manoeuvres, undermining their → autonomy and affecting their dignity. Lack of information also renders patients' freedom to decide on the treatment meaningless. Finally,

- **Who may be concerned?**

- When purchasing, procuring, designing (e.g. in-house developments) and using AI systems, **hospitals and other healthcare settings** as well as **physicians** should pay sufficient attention to aspects of understandability (→ understandable explanations, → explanations of AI systems and their outcomes), rooted in technical → interpretability and explainability of the system and its → outputs and output data.
- **AI developers** should generally keep → intelligibility and the need for medical justifiability in mind when choosing algorithms. That means, where feasible, to give preference to AI models that are more readily intelligible ('inherently intelligible' models; [Weld & Bansal, 2019](#)) and interpretable to physicians and patients over (per-trained) neural networks that may be readily available, apparently highly performant but are typically *a priori* inscrutable (see comments by [Rudin, 2019](#), [Afnan et al., 2021](#); [Garrett & Rudin, 2023](#); [Vokinger et al., 2019](#)).
- When using black-box AI (e.g. in medical diagnostics), **AI developers** should, already during the design phases, plan for the use of appropriate "explainable AI techniques" (XAI), that is the use of a second (post-hoc) model to explain the actual model's behaviour. This is widely propagated to support explainability, although from an epistemological point of view there are various issues with the assumption that this renders true 'explanations' (see [Rudin, 2019](#)). In any case, explainability may support → intelligibility with the ultimate aim of constructing complete → understandable explanations of AI systems and their outcomes (→ explanations of AI systems and their outcomes) that are characterised by allowing to *predict* properties, outputs, forecasts etc. The property of predictability (or lack thereof) can serve as a means of interrogating the possible validity of a scientific explanations including relevant underlying observations, rules, laws (formalised as sentences). Alternatively, a certain degree of inherent intelligibility may be built into black box models by → causal machine learning or → neurosymbolic AI.

- **AI developers** also need to pay attention that explanations of AI outcomes need to be targeted to various groups: clinicians may require more technical information, whereas information for patients should be understandable to the non-medical expert, with options for interested patients to access more detailed information. Ideally, commercial developers, healthcare professionals and patient associations should interact in order to develop processes and good practices for intelligible and understandable explanations of AI system's outcomes used in medicine and healthcare.

## Right to know and right not to know

**Parent term:** Dignity, freedom & autonomy

The Oviedo Convention ([Council of Europe, 1997a](#), see → DIGNITY, FREEDOM AND AUTONOMY) outlines a “right not to know” (RNTK), i.e. the right of patients not to know pertinent medical information concerning their own health (Article 10 point 2). RNTK is also mentioned in the Lisbon declaration of 1981 of the World Medical Association ([WMA, 1981](#)).

Nevertheless, the “right not to know”, is subject to ongoing debate in medicine ([Davies & Savulescu, 2020](#)). It can, in any case, be considered a *desideratum* for a proportion of patients, who wish to live without the burden of knowledge about the precise condition and, in particular, its prognosis.

In the context of the use of AI in healthcare coupled to the increasing shift towards home-based care ([WHO, 2021](#)), it is important that such systems are designed in a manner as to enable patients to exercise this right, e.g. by making provision that specific data or results of a monitoring device are not visible or accessible to the patient- provided the patient made explicit use of his/her/their right not to know.

- **Who may be concerned?**

*AI developers and healthcare institutions and other providers, in particular when involved in telemedicine using AI-enabled medical device software that collects health information of patients.*

## Right to know if AI system employed

**Parent term:** Dignity, freedom & autonomy

For reasons of maintaining the dignity of patients, for patients trust in healthcare and health systems as well as in regard to → upholding a trustful patient-physician relationship, it is important that patients are made fully aware when non-human actors are used in healthcare contexts ([European Commission - European Group on Ethics in Science and New Technologies, 2018](#)).

This right holds for both cases:

- a) **situations where patients directly interact with AI**, e.g. → conversational agents (in particular where such systems are sufficiently sophisticated to appear like human actors; → affective computing) and
- b) **situations where results of AI systems are used in the context of healthcare**, e.g. for diagnosis or prognostic purposes which typically have an impact on consequential decision making (e.g. treatment choices, treatment planning).

The right to know if an AI system is employed is not only an ethical issue of → DIGNITY, FREEDOM & AUTONOMY but obviously also one of → TRANSPARENCY (→ Transparency of human-AI interaction).

- **Who may be concerned?**

Healthcare settings (e.g. hospitals), physicians, clinicians, healthcare systems (e.g. when using → conversational agents).

## Medical privacy / health privacy

**Parent term:** Dignity, freedom & autonomy

The **Oviedo Convention** ([Council of Europe, 1997a](#); see → DIGNITY, FREEDOM AND AUTONOMY) outlines under Article 10 point 2 that “*Everyone is entitled to know any information collected about his or her health.*” The **Convention for the protection of individuals with regard to automatic processing of personal data**. ETS No. 108 and its modernising protocol of 2018 ([Council of Europe, 1981; 2018](#)) set out a series of rights in relation to medical / health privacy that are particularly relevant for the use of AI in medicine and healthcare. These are in Chapter II on basic principles and Chapter III on trans-border flow of data. Many of these rights mirror protections in relevant legislations and *vice versa*, e.g. the EU’s General data protection Regulation (GDPR; [EU 2016](#)) (Indeed the modernisation of the convention and development of EU law happened in close collaboration). These rights of patients result in relevant obligations for those actors and organisations that process data in the context of AI development and use. For translational concepts see → PRIVACY PROTECTION.

Importantly, the advent of AI and the **shift from stationary care situations towards home care** ([WHO, 2021](#); Chapter 3.1 - Emerging trends in the use of AI in clinical care), the **use of wearable medical devices connected to other devices via the internet of things** (IoT) will likely increase the complexities around technological needs and vulnerabilities ([Rami et al., 2023](#)), **including appropriate steps for patient consent management** given the increasingly complex value chain of various actors and distributed responsibilities ([Mittelstadt, 2021](#)).

- **Who may be concerned?**

AI developers (e.g. when collecting health data for model training), healthcare institutions and other providers, clinicians, physicians, health information professionals.

## Consent concerning collection of personal data / medical information

**Parent terms:** Dignity, freedom & autonomy – Medical privacy / health privacy

Consent regarding the collection of personal data / medical information is a key principle of → data privacy.

Article 5 (Legitimacy of data processing and quality of data) of the Convention for the protection of individuals with regard to automatic processing of → personal data ([Council of Europe, 2018; amended 2018](#)) sets out fundamental rights in regard to data processing:

1. *Data processing shall be proportionate in relation to the legitimate purpose pursued and reflect at all stages of the processing a fair balance between all interests concerned, whether public or private, and the rights and freedoms at stake.*
2. *Each Party shall provide that data processing can be carried out on the basis of the free, specific, informed and unambiguous consent of the data subject or of some other legitimate basis laid down by law.*

The Oviedo Convention (Council of Europe, 1997a; see → DIGNITY, FREEDOM AND AUTONOMY) correspondingly stipulates the right to privacy concerning **personal data / medical information** (Article 10):

*“Everyone has the right to respect for private life in relation to information about his or her health. Everyone is entitled to know any information collected about his or her health.”*

It is therefore fundamental that patients are aware and, if necessary, are made aware that they require to **provide their consent** in cases their personal data including medical information is collected. These rights include the right to **object to such collection**.

This concerns various scenarios and purposes for collecting data, for instance

- electronic health records for facilitated patient management,
- for purposes of training AI systems,
- for temporarily storing → post-deployment input data (medical images or textual information relating to patients) on local hospital servers or through means of other IT infrastructure (e.g. cloud-based) using, where necessary relevant, specific enabling technologies (for an ontology of value chain enablers and technologies: [Reina & Griesinger, 2024b](#)).
- the storage of personal / medical data for training healthcare professionals locally in their hospital.

These provisions are reflected in relevant legislations: in the EU, the consent concerning the processing of personal data and data access are regulated via the General Data Protection Regulation and the EU's Data Act (→ **PRIVACY PROTECTION**). Finally, electronic information handling and, where necessary, information exchange (e.g. upload on cloud server) may come with cybersecurity risks in case vulnerabilities are not appropriately managed (→ **PRIVACY PROTECTION**; → **NON-MALEFICENCE – Risks related to insufficient robustness / resilience**).

- **Who may be concerned?**

*AI developers (e.g. when collecting health data for model training), healthcare institutions and other providers, clinicians, physicians, health information professionals.*

### Right of intelligible information about personal data being processed

**Parent terms:** Dignity, freedom & autonomy – Medical privacy / health privacy

Article 9 of the Convention for the protection of individuals regarding automatic processing of personal data ([Council of Europe, 1981; amended 2018](#)) sets out fundamental rights of the “data subject” (i.e. the person whose data are processed). Points b and c stipulate that data subject have the right:

- „**b. to obtain, on request, at reasonable intervals** and without excessive delay or expense, **confirmation of the processing of personal data relating to him or her**, the **communication in an intelligible form of the data processed**, all available information on their origin, on the preservation period as well as any other information that the controller is required to provide in order to ensure the transparency of processing in accordance with Article 8, paragraph 1;
- c. to obtain, on request, knowledge of the reasoning underlying data processing** where the results of such processing are applied to him or her.

These rights are enshrined in relevant legislation, e.g. the **EU's General Data Protection Regulation** (Article 15, Recitals 63 and 64; [EU, 2016b](#)) which outlines that **data subjects (individuals whose data are being processed) have the right to request a copy of any of their personal data which are being ‘processed’** (i.e. used in any way) by ‘controllers’ (i.e. those who decide how and why data are processed), as well as other relevant information. In the literature, this right is often referred to as ‘data subject access requests’, or ‘access requests’.

- **Who may be concerned?**

*AI developers (e.g. when collecting health data for model training), healthcare institutions and other providers, clinicians, physicians, health information professionals.*

### Literature and further reading:

- Irish data Protection Commission (2024) provides an informative website on GDPR provisions:

Right to demand rectification or erasure of data processed not in line with relevant provisions

**Parent terms:** Dignity, freedom & autonomy – Medical privacy / health privacy

Article 9 of the Convention for the protection of individuals with regard to automatic processing of personal data ([Council of Europe, 1981; amended 2018](#)) sets out fundamental rights of the “data subject” (i.e. the person whose data are processed). Point e stipulate that data subject have the right:

*“to obtain, on request, free of charge and with- out excessive delay, **rectification or erasure**, as the case may be, **of such data if these are being, or have been, processed contrary to the provisions of this Convention”***

This right is reflected in relevant legislations, e.g. the EU’s GDPR (Article 16 and 19) which enable to demand rectification (in case of inaccurate data) or data completion (in case of incomplete data).

- **Who may be concerned?**

*AI developers (e.g. when collecting health data for model training), healthcare institutions and other providers, clinicians, physicians, health information professionals.*

## A.4 Privacy protection

### Concept description

Privacy protection captures the obligation of organisations or actors processing personal data to protect the privacy of the natural persons from which these data have been obtained. The obligation of privacy protection is a result of the right of individuals for privacy protection - a fundamental human right. In this section, we address the **obligations** of those that process personal data, while the **rights** of individuals are addressed under the ethical principle of → DIGNITY, FREEDOM AND AUTONOMY.

As a fundamental human right, privacy protection is laid down in relevant treaties and charters in Europe and the EU:

- The **convention for the protection of individuals with regard to automatic processing of personal data** ([Council of Europe, 1981](#)) outlines relevant principles.
- Of particular relevance in the context of health and medicine is the **Oviedo Convention on human rights and biomedicine** (ETS No. 164) by the Council of Europe ([Council of Europe, 1997](#)), requiring the respect for private life in regard to medical information ("medical privacy"), including the freedom to give or withhold consent in regard to the collection, processing, storage and use of information about one's health (e.g. for electronic health record and/or biomedical research). Cybersecurity plays an important role for ensuring privacy of data (→ risks related to insufficient robustness / resilience).
- In the EU, the **Charter of Fundamental Rights of the European Union** ([EU, 2000](#); see Article 8(1)) and the **Treaty on the Functioning of the European Union** (TFEU; [EU, 2007](#)) provide that everyone has the right to the protection of their personal data.
- The **EU general data protection regulation** (EU GDPR; [EU, 2016](#)) provides a legal framework for data protection in the EU with impacts on data-related services globally.

Technological developments may present novel challenges also for data privacy. Reports on health data governance ([OECD, 2016](#)) and on data privacy, ethics and protection in view of big data use ([United Nations, 2017](#)) have been published. In 2024 the OECD published a further report on AI, data governance and privacy ([OECD, 2024e](#)), with a focus on ways in which policy communities can work together to address related risks, especially those associated with the **emergence of generative AI** (→ see foundation models; → generative AI).

### Explanatory note

We structure privacy protection according to two fundamental aspects: **data protection** and **data security**.

Under **data protection** we address:

- **The lawful, legitimate and fair/ethical use of data:** This includes purpose specification, use limitation, proportionality of data use in relation to the need (of the project) as well as data retention and minimization ([UN development group, 2017](#)).
- **A forward-looking approach regarding risks for privacy protection:** risk identification, data protection and benefits. This can (and in some jurisdictions) must imply relevant impact assessments, notably the data protection impact assessment (DPIA) and the data transfer impact assessment (DTIA) under the EU's GDPR ([EU, 2016](#)).
- **Rights of access, rectification, erasure:** Management of access rights of individuals to their data, but also processes for the rectification or erasure of data that were not processed in line with relevant provisions
- **Governance and transparency:** Data governance, accountability and transparency measures.

- **Privacy-preserving techniques** for model training

Under **data security** we address:

- Considerations and mechanisms for preventing **unauthorised access** to personal data and
- **Information security**, using the well-established concept of the preservation of *confidentiality, integrity and availability* ('CIA') of data / information (e.g. metadata, statistics of actual data)

The concepts discussed in both sections, data protection and data security, concern **all data categories in the context of AI applications in the health domain** such as → **training data**, → **testing data**, → **post-deployment input data** (e.g. radiological images, electronic health records - patient data, medical history, diagnoses, medications, treatment plans, clinical decision support data etc.), big health data, public health data, clinical trial data, **patient-generated data** (e.g. data from **wearable devices**).

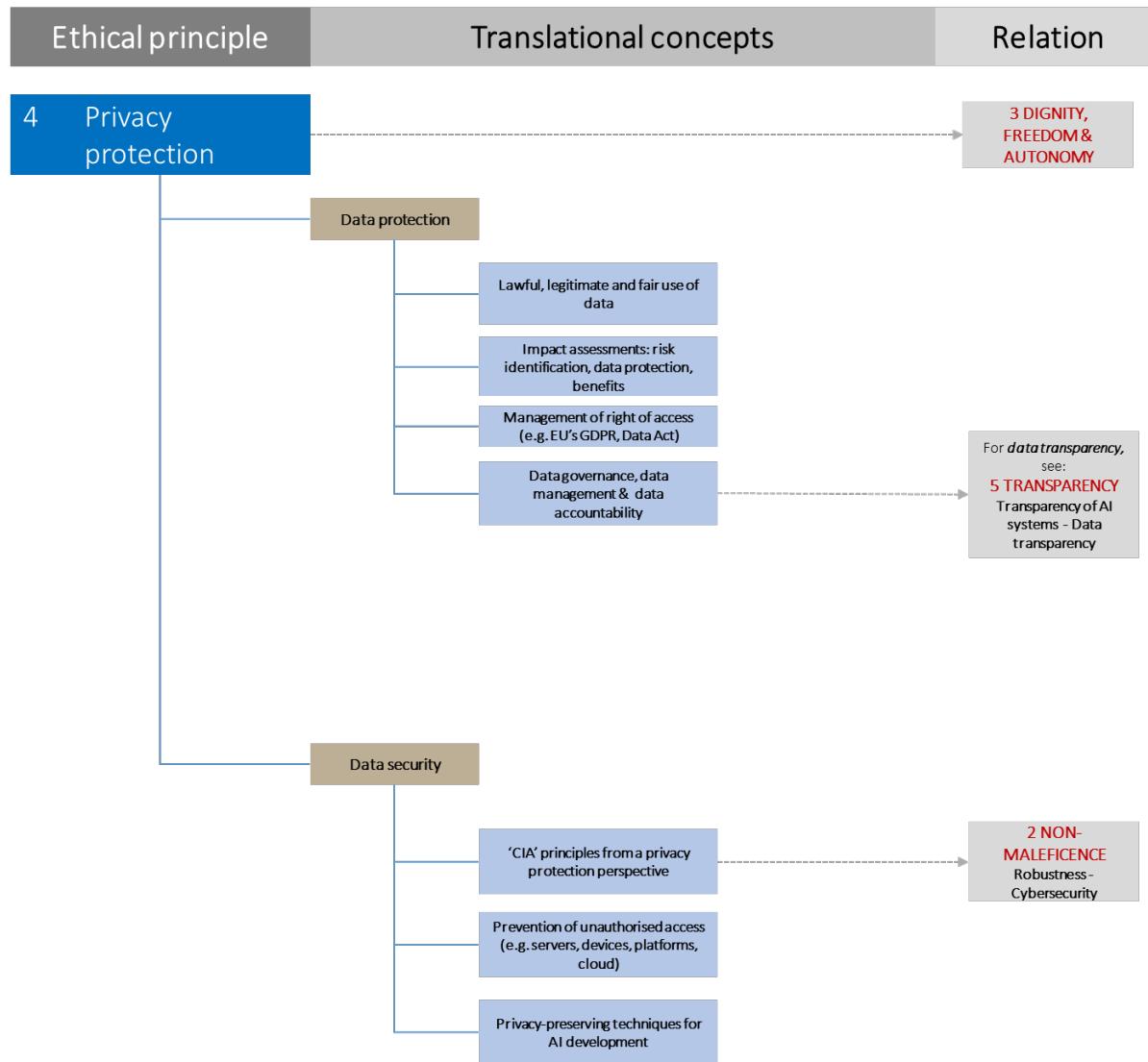
Aspects of transparency of data are addressed under the ethical principle of → **TRANSPARENCY** (→ **transparency of AI systems: evidence**).

#### Term relationship

Related terms:

- **DIGNITY, FREEDOM AND AUTONOMY**
- **NON-MALEFICENCE**
- **RESPONSIBILITY**
- **TRANSPARENCY**

Ontological organisation of the ethical principle of “privacy protection” and its translational concepts. Relations to other translational concepts are indicated in grey / stippled arrows.



## Privacy protection: translational concepts

### Data protection

#### Parent term: Privacy protection

Data protection focuses on the legal, administrative, organisational and ethical aspects of handling data of natural persons, ensuring that individuals' rights and freedoms are respected. This notably includes the right to privacy of medical and health information (→ [medical privacy / health privacy](#)). A recent study found that data privacy, among trust, is one of the major obstacles for AI uptake in healthcare ([Ahmed et al., 2023](#)) and for delivering clinical impact ([Kelly et al., 2019](#)). It has been identified as a major ethical concern in relation to AI use in medical imaging ([Brady & Neri, 2020](#)).

- For AI systems in healthcare, patients must be asked for their **consent in case their data are collected, stored, processed and used.**
- Patients must be informed on **how their data are being used** (i.e. for what purpose), including in case their data are shared with AI developers (e.g. local hospital research teams or companies) for purposes of **training/developing new AI models** (e.g. for health research, healthcare), but also for **purposes of training of medical students or healthcare professionals**.
- Patients should be made aware about the **eventual fate of their data**, including protocols for data destruction, including in case of the → [decommissioning / retirement](#) of AI system's. This includes AI systems under evaluation, e.g. in the context of → [model evaluation exercises](#) or → [clinical investigations](#) prior to conformity assessment and/or authorization.
- Patients need to be **informed of their legal rights to access their own data** and what steps are taken to ensure privacy protection and → [data security](#).

The following translational concepts will go into more details.

### Lawful, legitimate and fair use of data

#### Parent terms: Privacy protection – Data protection

Key principles regarding the lawful, legitimate and fair data use must be considered, inter alia:

Data must be collected, analysed and otherwise processed "*through lawful, legitimate and fair means*" ([United Nations, 2017](#)), irrespective whether the data are done directly or through collaboration with a third-party data provider.

- **Legal requirements:** Applicable laws must be identified concerning data protection and privacy (e.g. EU's [GDPR](#); [EU, 2016](#)).
- **Ethical requirements:** Data use must be fair, not violate human rights, must consider the legitimate interests of the persons of whom the data have been obtained.
- **Consent:** Individuals whose data are collected and processed must have given consent to the relevant data processing activities; this is of particular sensitivity in the area of health since health data are both, personal and highly sensitive (see also → [DIGNITY, FREEDOM AND AUTONOMY](#))
- **Purpose specification:** A clear purpose specification must be framed concerning the use of the data and this purpose cannot be changed post hoc, unless there is a legitimate basis ([UN, 2017](#)). It should be as narrowly and precisely defined as possible.
- **Proportionality of data** use in regard to specified goal (reflected in the purpose specification) must be ensured.

Many of these provisions are outlined in other relevant conventions (e.g. Convention for the protection of individuals regarding automatic processing of personal data. ETS No. 108 ([Council of Europe, 1981](#), Modernised convention of 2018) and reflected in legislations such as the EU's General Data Protection Regulation (GDPR) introduced in 2016 already. The GDPR is one of the most complete data protection legislations globally.

For additional references, see next concept:

→ Impact assessments: risk identification, data protection, benefits

- **Who may be concerned?** Companies and healthcare settings (e.g. hospitals) developing and using AI systems, health researchers, healthcare organisations and healthcare system managers.

## Impact assessments: risk identification, data protection, benefits

**Parent terms:** Privacy protection – Data protection

In order to identify, assess and anticipate risks associated with possible impact of project, programs or systems on privacy and data protection, organisations or actors need to conduct relevant impact assessments, of which some are legally required under specific jurisdictions.

- **Privacy impact assessment (PIA):** is a systematic approach for identifying and assessing possible risks resulting from activities (program, system, and project) and relating to privacy resulting from the collection, use, handling, and disclosure of personal information. PIA may also include management and mitigation strategies concerning identified risks. PIA can be seen as a component of a wider risk management approach, including cybersecurity-related risks (see → [Data security](#)) (see also [Mollaefar & Ranise, 2023](#); [ComplianceAspekte, 2024](#)).
- **Data protection impact assessment (DPIA):** legally mandated by the EU's General Data Protection Regulation (GDPR) (see Art. 35, [EU \(2016\)](#)) under specific circumstances, such as using new technologies and where the processing brings a high risk to the rights and freedoms of a natural person. See also [GDPR.EU, 2025](#).
- **Data transfer impact assessment (DTIA):** this type of assessment might be requested when there is the need to transfer personal data to a third country or an international organisation (e.g. Art. 46 GDPR). For trans-border flows of personal data, see also [Council of Europe \(1981, modernised in 2018\)](#).

- **Who may be concerned?** Companies and healthcare settings (e.g. hospitals) developing and using AI systems, health researchers, healthcare organisations and healthcare system managers.

## Management of right of access (e.g. EU's GDPR, Data Act)

**Parent terms:** Privacy protection – Data protection

The users' right of getting adequate information on the data being processed play a central role in relevant conventions (e.g. Convention for the protection of individuals with regard to automatic processing of personal data, [Council of Europe, 1997](#)). In the EU, a data protection framework consisting of different legislations regulates various aspects in relation to data processing, data use and information rights.

Of particular relevance for data use in the context of health and medicine (e.g. for AI tools, AI research, health research etc.) are two EU laws that respectively regulate citizens' fundamental right to data protection and enhance the EU's data economy: the [GDPR \(EU, 2016\)](#) and the [Data Act \(EU, 2022\)](#).

The **EU GDPR** stipulates that the data subjects have the rights to obtain access to a copy of his/her/their personal data being processed (Art. 15 GDPR, Recitals 63 and 64). This “access” to personal data is an obligation of the data controller. The modalities to exercise the right of access as stipulated in Art. 15 GDPR, are described in [ENISA, 2022](#). If **personal data is transmitted to a third country** without an adequate level of protection, data subjects must be informed of all appropriate safeguards which have been taken.

The right of access includes the following aspects:

- the purposes of the data processing (→ lawful, legitimate and fair use of data)
- the categories of personal data concerned
- the recipients or categories of recipient to whom the personal data have been or will be disclosed
- where possible, the envisaged period for which the personal data will be stored, or, if not possible, the criteria used to determine that period
- the existence of the right to request from the data controller rectification or erasure of personal data or restriction of processing of personal data concerning the data subject or to object to such processing
- the right to lodge a complaint with a supervisory authority (see also Oviedo Convention ([Council of Europe, 1997](#)) in regard to rights to request erasure or correction of data)
- where the personal data are not collected from the data subject, any available information as to their source
- the existence of automated decision-making

The **EU Data Act** is consistent with the existing rules on the processing of personal data under the GDPR (see [European Commission, 2025b](#)). It provides for additional elements, relevant in the current context of health and medicine. The Data Act for instance states that **products connected to the internet (such as medical or health devices (see Recital 14))**, e.g. wearable devices monitoring blood pressure, heart rate, blood sugar or other physiological parameters) shall be designed and manufactured, and related services shall be provided, in a manner so as to **ensure that data generated by their use are directly accessible to the user** (Data Act, Article 3).

If data cannot be made directly accessible, it must be made available on request without undue delay (Art. 4 Data Act) by the data holder (e.g. the company that makes the connected product or that provides a related service). If the user wishes to share this data with another entity or individual ('third party'), they can either do so directly or they can ask the data holder to share it with a third party of their choice (Art. 5 Data Act).

- **Who may be concerned? All stakeholders along the AI development and use pathway, in particular AI developers, companies and healthcare settings, health system organisations, health researchers, healthcare system managers and patients.**

**Parent terms:** Privacy protection – Data protection

The term 'data governance' has slightly different meanings depending on the broader user and policy context:

**1) Processes to manage data assets across the data life cycle**

Generally, **data governance can be understood as a framework of necessary processes and procedures set-up at organisations** (e.g. companies, public organisations and regulatory bodies etc.) **and entailing a comprehensive and systematic approach to overseeing an organisation's data assets**, with a primary focus on data quality throughout the data "lifecycle" based on data controls. [Google \(2024\)](#) for instance defines data governance as follows: "*Data governance is a principled approach to managing data during its life cycle, from acquisition to use to disposal.*" In this sense, data governance is related to data management. For a scoping overview of the development of **health data management in healthcare and research**, see [Ismail et al., 2020](#). See also the report on AI in processing and generating new data ([European Commission, 2024a](#)) and the guide to data governance for data privacy by [Data-Meaning \(2024\)](#).

Important elements of data governance are availability, usability, consistency, data integrity and data security, in particular through safeguarding sensitive information and upholding individual privacy rights. This involves establishing a robust framework that integrates privacy-centric policies, procedures, and standards to ensure the confidentiality, integrity, and availability of personal data (→ CIA principle from a privacy protection perspective).

**Accountability** for harm resulting from poor data quality or violations of legal frameworks and ethical principles are another element of data governance. To this end, **data governance requires continuous oversight and auditing to guarantee that data are accurate, consistent, and used in a manner that respects privacy principles and supports organisational objectives** while **minimizing risks and ensuring compliance** with relevant data protection regulations (e.g. EU's GDPR; [EU, 2016](#)). In the real-world, data governance may show elements of "authority multiplication" and "actor subordination" ([Paparova et al., 2023](#)).

Effective data governance requires:

- the definition of clear roles and responsibilities
- the implementation of data handling protocols and data controls
- the designation of data stewards who can monitor and enforce adherence to these protocols

**2) Mechanisms to ensure compliance with laws including concerning privacy and ethical integrity when dealing with data**

Data governance is also used for notions of ethical integrity and adherence to and compliance with relevant law. For instance, the UN Development Group in its guidance on data privacy, ethics and protection ([UN; 2017](#)) outlines: "*Appropriate governance and accountability mechanisms should be established to monitor compliance with relevant law, including privacy laws and the highest standards of confidentiality, moral and ethical conduct with regard to data use.*" Further, assessments of risks, harms and benefits should be part of a data governance framework in relation to specific use of data.

**3) A management framework for training, validation and testing data**

In the EU's AI Act, data governance (Article 10; [EU, 2024a](#)) focuses on the governance and management of data used for machine learning, i.e. training, validation and testing data used for the development of AI systems. Article 10 outlines general practices of data governance and management (Paragraph 2; see also: [European Commission, 2024](#): page 46), general requirements of data quality, relevance, representativeness, completeness, correctness, statistical properties etc. (Paragraph 3), context-dependent aspects (Paragraph 4), the processing of data required to tackle bias detection/monitoring (Paragraph 5). Concerning this point, see also → FAIRNESS – Monitoring and mitigation of possible discrimination throughout the evidence pathway.

#### **4) Rules and means to use data through sharing mechanisms, agreements and technical standards**

In the context of data sharing, the European Commission, in the context of a proposal on “data solidarity” defines data governance in that specific context as follows: *“Data governance refers to a set of rules and means to use data, for example through sharing mechanisms, agreements and technical standards. It implies structures and processes to share data in a secure manner, including through trusted third parties.”* ([EU Commission – Regulation on data governance, 2020; see also WHO, 2021a, page 83, section: Evolving approaches to consent](#)).

- For **data transparency**, see → **TRANSPARENCY**.

- **Who may be concerned?** Companies and healthcare settings, health system organisations, health researchers, healthcare system managers.

## Data security

### **Parent term:** Privacy protection

Data security refers to the measures put in place to protect data from unauthorized access, use, disclosure, disruption, modification, or deletion. These elements can be captured under data confidentiality, integrity, and availability, i.e. the so-called data ‘CIA model’ (→ CIA principle from a privacy protection perspective) ([Unitrends, 2024](#)).

Typically, data security involves technical measures, controls and protocols used to protect data from unauthorized access, data breaches and other cybersecurity threats.

Nevertheless, it is essential that the measures taken to protect private information and to keep data safe, do not unduly interfere with the *purpose* of legitimately collected data (→ lawful, legitimate and fair use of data).

### CIA principle from a privacy protection perspective

#### **Parent terms:** Privacy protection – Data security

The → CIA principles can be used as a useful basic framework by organisations that handle sensitive information, including those in the healthcare sector. By implementing the CIA principles, businesses can safeguard their information systems against unauthorized access, theft, or manipulation of sensitive data, thereby upholding the trust of their users and protecting their reputation. The three components of the CIA principle (→ CIA principle from a privacy protection perspective) contribute to ensuring that patients’ data remains secure, accurate, and accessible only to authorized individuals or processes.

The CIA principle (or CIA triad) can be complemented by other components to complete it in view of data privacy. Hansen et al have proposed six fundamental “protection goals for privacy engineering” and schematically summarised this in a six-pointed star of three axes, where each axe represents one pair of opposing protection goals ([Hansen et al. 2016](#)): confidentiality versus availability, integrity versus intervenability, and transparency versus unlinkability. According to ([Hansen et al. 2016](#)), intervenability is a way of ensuring that data subjects have the ability to control how their data is processed and by whom. This property is discussed in → management of right of access. Hansen et al. use the term ‘unlinkability’ in the context of “*data avoidance*”: this means that data should only be collected, processed, and stored when necessary. This topic is discussed in → lawful, legitimate and fair use of data. For data transparency, see → **TRANSPARENCY**. The proposals of Hansen (of ULD in Germany; [Hansen et al., 2016](#)) and NIST in the USA ([NIST, 2017](#)) have been reviewed by Covert and colleagues ([Covert et al. 2020](#)).

In any case, in the context of AI-powered health applications, the CIA principle takes on particular significance:

- confidentiality ensures the privacy of patient data throughout the AI development and deployment process
- integrity guarantees the accuracy and reliability of the data as it flows through the AI pipeline
- availability ensures that authorized personnel have timely and uninterrupted access to health data and AI workflows, regardless of their location or device

Even if preserving confidentiality often implies data privacy, [El Mestari et al. 2024](#) have alerted to the need to discern, in the context of → machine learning, *confidentiality from data privacy*. There is a difference in regard to what information the “sheer” or actual data (potential confidentiality issues) provide as opposed to what information may be retrieved by accessing *statistics about the data* (potential privacy issues). The authors propose more granular definitions for confidentiality and privacy:

- *Confidentiality of the data ensures that there is no explicit disclosure of the data or of certain parts of the data. In other words, the confidentiality of the data is preserved if the raw data have never been subject to unauthorised access or disclosure.*
- *Privacy of the data is protected when it is ensured that unauthorised actors cannot retrieve or reconstruct sensitive information, e.g. by accessing information exchanged during → federated learning & split learning, including from metadata or statistics about the data.*

#### Prevention of unauthorised access (e.g. servers, devices, platforms, cloud)

##### **Parent terms:** Privacy protection – Data security

An unauthorised access occurs when “*a person gains logical or physical access without permission to a network, system, application, data, or other resource*” ([NIST, 2004](#)). Preventing unauthorised access is paramount to ensure data security considering the great impact that a loss of data confidentiality can have on both organisations and individuals (see also the proposal concerning encryption by [Van Daalen, 2023](#)).

A useful high-level architecture and a set of capabilities that can be used to identify and protect assets from unauthorized access and disclosure is outlined NIST’s Computer Security Incident Handling Guide (Chapter 4 Architecture, [NIST 2024](#)):

“*Each of the capabilities plays a role in mitigating data confidentiality attacks:*

- **Data Management** allows discovery and tracking of files throughout the enterprise.
- **Data Protection** involves encryption and protection against disclosure of sensitive files.
- **Access Controls** allows organisations to enforce access control policies, ensuring that only authorized users have access to sensitive files.
- **Browser Isolation** protects endpoints in the organisation from malicious web-based malware by sandboxing and containing executables downloaded from the internet.
- **Policy Enforcement** ensures that endpoints in the organisation conform to specified security policies, which can include certificate verification, installed programs, and machine posture.
- **Logging** creates a baseline of a normal enterprise activity for comparison in the event of a data confidentiality event.
- **Network Protection** ensures that hosts on the network only communicate in allowed ways, preventing side-channel attacks and attacks that rely on direct communication between hosts. Furthermore, it protects against potentially malicious hosts joining or observing traffic (encrypted or decrypted) traversing the network.”

A similar list of security measures that establish and maintain appropriate (logical) access controls for the network and information systems, to prevent unauthorized access, modification, or deletion of data is presented in ENISA'S Technical Guideline on Security measures ([ENISA, 2014](#)) and Data Protection Engineering ([ENISA, 2022](#)).

In this context, encryption technologies, that are a subset of cryptographic technologies ([Van Daalen, 2023](#)), play an important role in reducing the risk of unlawful access to information. Encryption technologies are for example involved in protecting information from unauthorised users, thus preserving confidentiality.

In parallel with encryption, mechanisms of de-identification (anonymisation or pseudo-anonymisation) of personal (health) data might be implemented in order to minimize risks in case of situations of unauthorised access to personal data.

Along with logical access controls, it is equally important to set up physical controls to protect network and information systems and facilities from **unauthorized physical access** ([NordLayer, 2024](#)) such as locks, guards, and access control cards, systems for biometric access control, surveillance cameras and intrusion detection sensors.

Finally, there may be specific considerations depending on medical field, e.g. radiological imaging (e.g. [Shah et al., 2023](#)).

## Privacy-preserving techniques for AI development

### **Parent terms:** Privacy protection – Data protection

There are various techniques and practices that can be adopted to safeguard patient privacy during the training and deployment of AI-based applications in healthcare. [Hagendorff, 2020](#) refers to these techniques as *privacy-friendly* and includes, for instance, cryptography and differential or stochastic privacy. A taxonomy of privacy-preservation techniques, focusing on healthcare domain applications, is presented in [Khalid et al. 2024](#) where the techniques are grouped in 4 groups:

1. Cryptographic techniques (e.g. homomorphic encryption, Secure Multiparty Computation)
2. Non-cryptographic techniques (e.g. differential privacy)
3. Decentralised systems (e.g. blockchain)
4. Hybrid privacy-preserving techniques (e.g. → **federated learning & split learning**)

A different approach is used in [El Mestari et al. 2024](#), where existing privacy preserving techniques are categorised according to phases of the machine learning pipeline from data preparation to model inference:

- 1) Mitigation techniques during the data preparation phase (e.g. perturbation techniques)
- 2) Mitigation techniques during the model building phase, e.g. private aggregation of teacher ensembles – PATE ([Papernot et al., 2018](#))
- 3) Mitigation techniques during the model serving phase (e.g. trusted execution environment)

Similar techniques are also outline in [Tran et al. 2024](#) and in [ENISA, 2022](#).

- **Who may be concerned?**

*Commercial AI developers including healthcare institutions, clinicians, physicians, health information professionals engaged in AI development for in-house use.*

## A.5 Transparency

### Concept description

Transparency is a central precondition for trust. While trust is a belief to start with, trust, to be maintained nevertheless requires evidence that is a) available and b) of sufficient quality and clarity (=transparent) to support the trustor's confidence in the trustee (someone or something). Only such evidence will support the trustee's 'trustworthiness' (→ TRUST AND TRUSTWORTHINESS). Thus, whenever evidence on relevant notions of trustworthiness (e.g. truthfulness, performance, safety, quality, equality, fairness) is not sufficiently transparent in the above meaning, the evidence is functionally non-existent to the trustor, leaving the trustor-trustee relationship (Mayer et al. 1995) dysfunctional.

It is not surprising hence that transparency is the most frequently used concept or principle (→ ethical principles) for AI (→AI principles and AI ethics guidelines) (Jobin et al., 2019). However considerable variation has been reported concerning the understanding of the dimensions of transparency, in particular the *domains* of transparency and *modes* of achieving it (Jobin et al., 2019; Ryan and Stahl, 2020; Kiseleva et al., 2022). In addition, there is perhaps insufficient discussion that transparency, for a variety of reasons, is not uniformly applicable to all contexts: it concerns various degrees of disclosure and detail depending on specific trustor-trustee communities (see also Rademakers et al., 2025). A critique of the 'transparency ideal' and its limitations have been made by Annany & Crawford, 2018.

In the following we briefly discuss 1) layers of trustor-trustee communities in regard to transparency and 2) outline briefly the various 'domains' of transparency, addressed as translational concepts in more detail. This includes also of transparency, including the type of evidence supporting notions of transparency.

#### 1) Communities of transparency

Transparency cannot be applied in a dichotomous way: i.e. either full transparency or no transparency. Full transparency (i.e. without any "secrets" see Alan Turing Institute, 2024) is neither realistic nor desirable. It would negatively impact innovation capacity and competitiveness of private organisations that require justifiable degrees of business secrecy.

Instead, depending on the community concerned, various degrees of transparent disclosure, communication and showing are required. This can be schematically depicted in a simple Venn diagram (Figure 13).

**Internal community** refers to the community that has access to fully transparent information of all aspects relating to the functioning and developing of an AI system, e.g. an organisation developing or deploying such a system. This will include information on specific approaches that may be protected by intellectual property rights and considered confidential business information (CBI). Protection of such information is essential for businesses to thrive and hence for competitiveness and innovation. Internal transparency should involve a maximum level of transparent recording and logging of all relevant decisions (→ traceability), including those that led to the development of a specific model (→ algorithm-to-model transition). This will allow understanding root causes of failures and correcting these in a timely manner. Even within an organisation there may be various layers of transparency, e.g. different degrees of information accessible to management, internal compliance experts, data scientist, modellers versus people working in production or marketing etc.

**Community of compliance and HTA experts** refers to expert communities (e.g. at Competent Authorities, notified bodies or → health technology assessment bodies) that require information that developers, developing organisations or companies consider confidential. Such information may be required to assess compliance with essential legal and regulatory requirements (e.g. safety, performance, benefit-risk) and to make assessments of the overall added value of an AI system for a (national or regional) health system, based on clinical, scientific, economic and ethical considerations.

**Community of society, patients and users** refers to an even more restricted level of transparency available to the general public and necessary for the safe use of the AI system, e.g. → intended use, → instructions for use, → foreseeable misuse, sufficiently audience-targeted → understandable explana-

tions on how the system works and how it produces outputs (→ intelligibility; → interpretability and explainability). Relevant actors (e.g. AI developers) should be **transparent to the extent possible** but need also to guard their legitimate interests, e.g. in terms of confidential business information.

Technical transparency concepts such as ‘traceability’, may permeate all three community layers, albeit with increasingly restricted availability of information associated with traceability processes. For example, companies will not disclose the fine-tuning of → hyperparameter settings during → machine learning to the general public but may need to communicate these, under terms of confidentiality, to the community of compliance experts in case of safety problems with an AI system.

**Figure 13.** Schematic Venn diagram of various communities regarding transparency of information concerning AI in healthcare. The internal community (yellow, e.g. developers, businesses, organisations) will need to implement maximum transparency (e.g. to trace and understand potential failures). The community of compliance and HTA experts (blue) may need access to parts of confidential business information to assess agreement with legal and regulatory requirements and/or added value of an AI tool for health systems. The community of society in general, patients and users require transparent information on how to safely use an AI system and should be informed about problems and issues and how these were resolved.



Source: own production

## 2) Translational concepts under ‘transparency’

There is currently no agreement regarding domains or concepts underpinning transparency: there is no taxonomy, neither within a scientific field (e.g. data science) nor between fields (e.g. data science and medicine) (Kiseleva et al., 2022). Transparency has been proposed as a central umbrella concept for AI development and use; it has been proposed to encompass interpretability, explainability, communication, auditability, traceability / record keeping, information provision (disclosure), data governance/management and documentation (Kiseleva et al., 2022).

Based on information in the literature (e.g. Jobin et al., 2019; Ryan & Stahl, 2021, EU HLEG, 2019; Kiseleva et al., 2022) we propose in this ontology **five translational concepts** for addressing transparency. We distinguish:

- 1) **Transparency of organisations and actors providing / deploying AI systems.** This includes appropriate communication and disclosure (information provision).
- 2) **Transparency human AI interaction:** clear indication in cases AI systems are used as interlocutors with human beings (e.g. conversational agents / chatbots)
- 3) **Traceability:** record keeping including on → algorithm-to-model transition and traceability of important failures or errors (failure transparency), in particular where these have affected safety and health and how these were corrected.

#### 4) Transparency of AI systems: evidence needs (for documentation)

- Evidence on → algorithm-to-model transition, i.e. the assumptions and decisions made when creating a machine-learning model and AI system
- Evidence on → data including data provenance and quality
- Information about frameworks processes along the life cycle and across the value chain (e.g. → AI governance and → AI management frameworks, systems and processes for ensuring quality and the management of risks) as well as in regard to fundamental rights and algorithmic impacts (→ AI impact assessment (AI-IA), → Fundamental rights and algorithm impact assessments)
- Evidence on interoperability needs and → value chain elements
- Information on training requirements

#### 5) Intelligibility of the AI systems operations and, in particular how and why it produces specific outcomes as opposed to other conceivable ones (counterfactual reasoning). This includes the more granular concepts of → interpretability and explainability\*\*, → explanations of AI systems and their outcomes, → understandable explanations and → explicability\*\*\*.

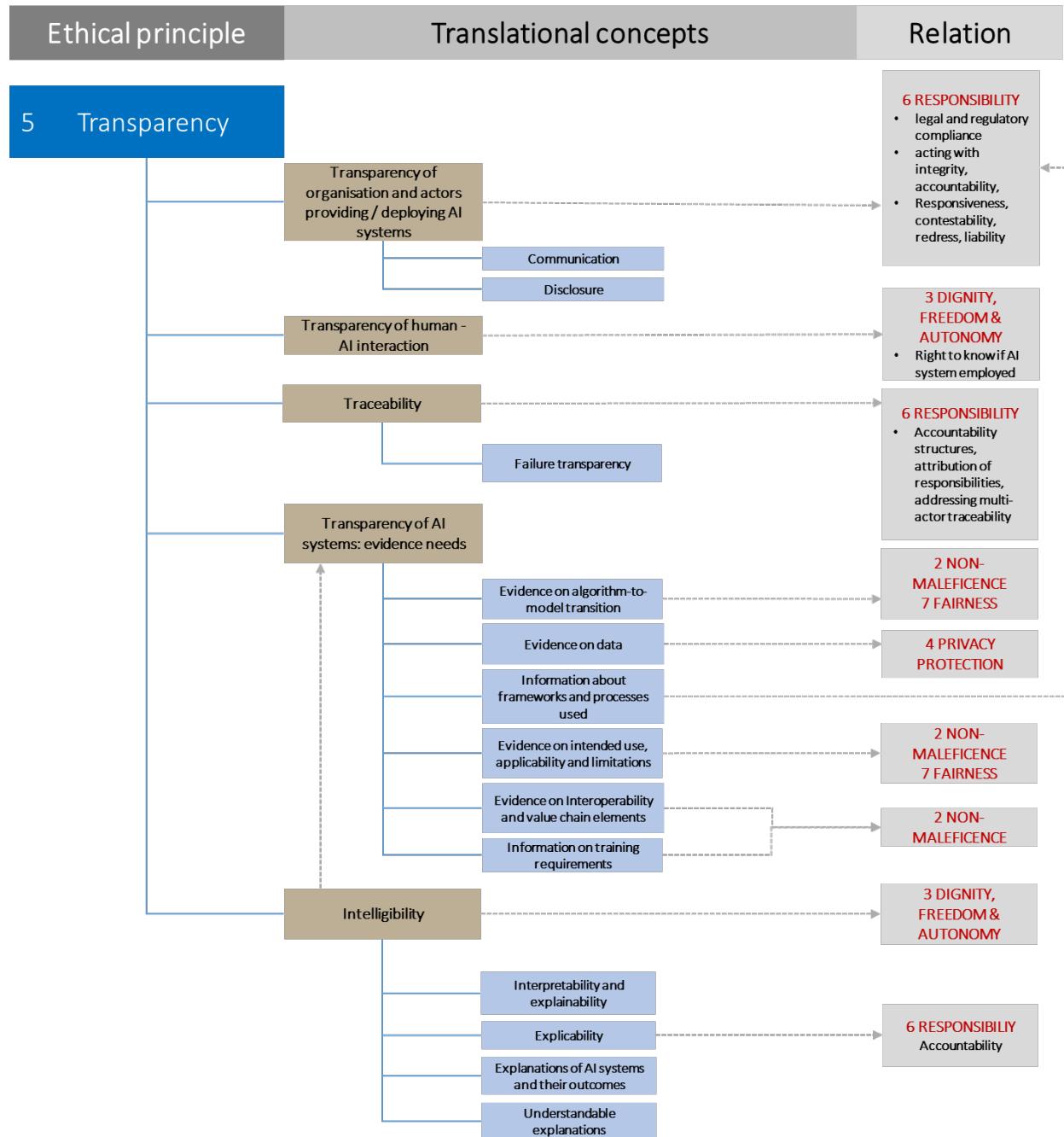
#### Explanatory note

\* In contrast to the proposal by [Kiseleva et al. \(2022\)](#), we list ‘accountability’ as a translational concept under → RESPONSIBILITY.

\*\*) *Interpretability* and *explainability* are often used near interchangeably or completely interchangeably (e.g. [Miller, 2019](#)). Some authors denote subtle differences in regard to the level and facility of comprehensibility of AI input-output relationships. Interpretability is used to denote an understanding of the input-output relationships of a model. Interpretability is used for all types of models, those that are inherently intelligible or understandable (white box or glass box models) but also models that are not immediately intelligible. Explainability refers to techniques for creating understanding on input-output relations of black-box models (e.g. [Weld & Bansal, 2019](#)), thus ultimately creating interpretability for such models. Black box models require (typically post-hoc) approaches of observing rules or ‘laws’ of decision making of the model and providing a deductive argument how these relate to or cause the phenomenon to be ‘explained’ (i.e. made plain / clear) – e.g. a specific outcome provided by an AI system. See → *interpretability and explainability*.

\*\*\*) *Explicability* should not be confused with either *explainability* or *interpretability*. Explicability is a term introduced by [Floridi et al., \(2019\)](#). Explicability combines notions of *organisational transparency* and *accountability* with a demand of *transparent and intelligible input-output functions*. For a comment, see [Herzog, 2022](#). Explicability is currently not widely used in the literature. It is however one of the four ethical principles of the guidance document of the European Commission’s independent high-level expert group on AI ([EU HLEG, 2019](#)).

Ontological organisation of the ethical principle of “transparency” and its translational concepts. Relations to other ethical principles are indicated.



## Transparency: translational concepts

### Transparency of organisations and actors providing/deploying AI systems

#### **Parent term:** Transparency

Relevant actors that develop and/or deploy AI systems ('providers' and 'deployers'; see → **actors as defined in the EU's AI Act**) should provide a minimum of transparency on their status and operations.

Relevant actors may comprise organisations, e.g. commercial enterprise/manufacturers, governmental bodies/agencies, NGOs, universities, university spin-offs, scientific consortia etc. as well as individual researchers that are engaged in modelling and/or development of AI systems that are made available to others (e.g. specific communities or the public).

Transparency considerations include:

- Relevant actors should be transparent about their legal statuses, their business model and their modes of distribution or deployment. This may include information on how they distribute and/or deploy an AI system (e.g. model platforms, commercial products).
- Relevant actors should consider using a published → **ethics code** that outlines how they approach ethical and trustworthy AI. This may include aspects of → **AI ethics** and → **AI principles and ethics guidelines**, including how these are practically applied for → **ethical evaluation of AI** and any relevant processes and decisions within the organisation
- Relevant actors may require disclosing how to approach, where applicable, → **AI impact assessment** and → **fundamental rights and algorithm impact assessments**.
- Where applicable, relevant actors should consider providing information on → **AI governance**, → **AI management** and → **AI risk management** frameworks and processes being used.
- Consideration should be given on disclosing how an organisation is tackling risks over the → **life cycle of AI in health** and across the → **value chain of AI**.
- Communication lines for receiving information from users, patients and other communities of society, e.g. concerning problems, failures, malfunctions (→ **incidents**, → **adverse events**), concerning → **usability** or issues related to drifts (→ **drift / shift in machine learning**).

### Communication

#### **Parent terms:** Transparency – Transparency of organisations and actors

#### **Provide communication**

AI actors should openly and transparently communicate relevant issues, in particular problems, failure and malfunctions (e.g. → **incidents** or → **adverse events**). For communication of problem-shooting or corrections of the AI system required in the post-deployment space, see also → **Correcting problems and failures including associated communication**.

Relevant actors should alert in particular healthcare users proactively to potential hazards and → **risks**, including → **biases** that were detected during post-deployment monitoring, post-market surveillance or audits (→ **auditability and auditing**).

#### **Receive communication**

Relevant actors should be open to receive communication from users and stakeholders by publishing communication lines for receiving stakeholder feedback. This includes situations of contestability and claims concerning redress and remedy (→ **contestability and challenge**; → **remedy and redress**).

## Disclosure

### **Parent terms:** Transparency – Transparency of organisations and actors

Both organisations that develop AI but also organisations that use AI (e.g. public bodies engaged in health system planning and management) should disclose as much information as possible about **how, why and for what purpose an AI system was developed, deployed and used**. This requires considering however legitimate interests of **confidentiality and intellectual property rights** or other justifiable reasons for confidentiality.

For example, organisations should disclose whether, to which extent and how often they use internal and external **auditing activities** including independent third-party auditing programmes in the context of, for instance, health data privacy protection measures (→ **PRIVACY PROTECTION**), anonymization, pseudonymisation, encryption.

Organisations (including healthcare systems) that use AI with a **potential high impact on society (including patients or specific patient groups)** should disclose information about

- the purpose of the AI system (e.g. for disease prediction, for public health surveillance, for research)
- what type of model is being used (e.g. → foundation models; → generative AI)
- what measures are in place to ensure → human agency (→ human oversight, → human primacy, → corrigibility) and dignity of patients and users (→ **DIGNITY, FREEDOM AND AUTONOMY**), in particular to ensure → patient primacy
- How key decisions during → algorithm-to-model transition were made and how these influence e.g. → model performance, → AI safety, → **PRIVACY PROTECTION** and → **FAIRNESS** considerations, including → bias mitigation.

Disclosure can be based on a → ethical evaluation of AI or a → AI algorithm impact assessment (AI-IA) (see for instance [Ada Lovelace Institute; 2022](#)).

Organisations and individuals that deploy AI systems for other purposes than healthcare (e.g. for research) should also enable independent → peer review and community discourse, with relevant discussions being publicly available, considering however intellectual property rights and necessary confidentiality.

Disclosure needs to take into account the recipient community (→ **TRANSPARENCY**), e.g. higher levels of disclosure are required for regulatory communities than for the general public.

## Transparency of human-AI interaction

### **Parent term:** Transparency

When using → **conversational agents** (chat-bots) in healthcare settings or health administrations (see also → **AI typology: interactive AI**), users must be made aware that they are interacting with an AI system ([EU HLEG, 2019; European Commission, 2021](#))

Interactive systems should not coerce, manipulate or condition human users. This includes AI systems intended to “nudge” patients towards healthier behaviours and/or preventive action (see → **AI and the development of healthcare**).

Furthermore, feedback communication lines should be available for gathering user experiences and problems when interacting with AI (→ **RESPONSIBILITY**).

This is not only a transparency matter, but also related to → **DIGNITY, FREEDOM AND AUTONOMY**; see → right to know if AI system employed.

## Traceability

### Parent term: Transparency

Traceability of an AI system refers to the possibility to trace back steps taken during development, production and deployment of an AI system through **identifiable record keeping and logging**. The OECD (2019a) has outlined the scope of traceability as follows: "Actors should ensure traceability, including in relation to datasets, processes and decisions made during the AI system lifecycle, to enable analysis of the AI system's outcomes and responses to inquiry, appropriate to the context and consistent with the state of art." Traceability and record-keeping have been emphasized also by the EU HLEG (2019).

Traceability allows to detect sources of failures, errors and harms (→ failure transparency) and is thus important for → NON-MALEFICENCE, → AI safety and the improvement of AI systems based on experiences, including those made in the post-deployment or post-market space. Traceability is closely linked to → algorithm-to-model transition (see also [FUTURE-AI, 2023](#)).

Important dimensions of traceability include:

- **Traceability and algorithm-to-model transition:** Traceability (proper record keeping) ensures that key *assumptions* and *decisions* (including their rationale) during the → algorithm-to-model transition can be traced back. This concerns for instance decisions on data sets and for modelling or when selecting algorithms or pretrained models. Traceability also concerns the processes that may need to be employed for ensuring quality and managing risks (→ AI risk management, → AI safety).
- **Traceability and intelligibility:** → Traceability relates to → interpretability and explainability, i.e. understanding the outcomes of an AI system and their impacts (i.e. the output data such as recommendations, classifications, decisions, actions used in a given context). This supports possible identification of the reasons why an AI outcome was erroneous (→ failure transparency) beyond expected margins (e.g. beyond measured → accuracy of a predictive system). This can help to prevent future mistakes and improve the design of the system.
- **Traceability and auditability:** Traceability supports auditability (→ accountability, → auditability). It thus supports trustworthiness (e.g. vis-a-vis inquiries, peer review and stakeholder discussion), enables compliance with regulatory requirements and supports responsibility and responsiveness. Traceability is critical also in the context of contested outcomes (→ RESPONSIBILITY – contestability and challenge, → remedy and redress, → liability).
- **Traceability and distributed responsibilities across the value chain and along the life cycle:** Finally, in situations where multiple actors are involved in developing and/or deploying an AI system and/or associated services, traceability may support understanding how harm originated ([Mittelstadt et al., 2016, 2021; Kiseleva et al., 2022](#)). Traceability should allow tracking decisions in regard to value chain elements and involved value chain agents as well as specific life cycle stages (e.g. use in real-world environments). Highly distributed responsibilities pose a challenge for traceability. A healthcare example is telemedicine for diabetes type I patients managing their condition with a wearable device connected to a hospital system. Other examples concern distributed responsibilities along the value chain, e.g. when data sets or pretrained models are purchased or otherwise acquired. Collaborative attribution of responsibilities and clear traceability rules across value chain and along the life cycle are important.

## Failure transparency

### Parent term: Transparency- Traceability

Failure transparency concerns the capability of ascertaining the reasons for an AI systems failure, in particular where this failure has led to harm ([Future of life institute, 2014a, b](#)).

Failure transparency relates to → TRANSPARENCY (notably → traceability), → intelligibility and to → NON-MALEFICENCE as well as → RESPONSIBILITY (e.g. accountability, liability) and

- **Failures that originate in the development process:** these may be traceable to assumptions made and decisions taken during → algorithm-to-model transition or to cybersecurity (e.g. data poisoning → robustness and cybersecurity: risks for value chain assets)
- **Failures due to insufficient verification, validation and evaluation activities** prior to deployment which may, for example, lead to undetected → bias with safety implications for (specific) patients or patient groups (see → risks related to pre-deployment evidence gaps)
- **Use failures:** Failures due to implementation in the → use environment (including technical interoperability), due to → use contexts that are out of the domain of applicability (→ applicability and limitations) or failures resulting from → foreseeable misuse or resulting from lack of training (see → training requirements).

Failure transparency is a prerequisite for being able to improve AI systems once deployed. It allows learning from errors, misuse and wrong outputs.

Failure transparency should be incorporated into processes of post-deployment monitoring and, in case of AI systems for healthcare, in → clinical evaluation (→ post-deployment monitoring; → post-market surveillance, market surveillance, corrective action).

## Transparency of AI systems

### **Parent term:** Transparency

A key element of the ethical principle of transparency is the availability of appropriately detailed **evidence** about the AI system. Relevant material (documentations, descriptions) should be available at the point of deployment so as to ensure → intelligibility and hence understandability of AI systems to different users (→ understandable explanations). Sufficient understandability is a necessary condition for informed consent by patients (→ ensuring the means for free and informed consent).

The evidence made available needs to strike a balance between **sufficient transparency for all stakeholders and legitimate rights to protect intellectual property rights and confidential business information**, so as to ensure competitiveness and innovativeness through protection of legitimate business interests.

Broadly, evidence on AI systems in the health domain should include:

- **How the AI system was developed** (→ algorithm to model transition, → transparency of AI systems: evidence – evidence on data, → transparency of organisations and actors providing/deploying AI systems)
- **Frameworks and processes used to ensure quality and safety as well as impact on fundamental rights.** This may include for instance i) → AI governance and → AI management frameworks and associated quality assurance systems and risk management frameworks, processes or systems (→ NON-MALEFICENCE); ii) processes related to → verification and → validation, iii) → AI impact assessment (AI-IA) or → Fundamental rights and algorithm impact assessments (where applicable).
- **What the AI system is intended to achieve, how it should be used and what are its limitations and use exclusions** (see →intended use, → applicability and limitations and, where needed, considerations of → use context, → use environment)
- **How the AI system should be deployed from a point of technical interoperability (where applicable).** This relates to → evidence on interoperability and value chain elements.
- **What training requirements are needed for the proper operation of the AI system,** ensuring safety (→ NON-MALEFICENCE), effectiveness (→efficacy, effectiveness and efficiency) and equal benefits (→ FAIRNESS)
- **How the AI system generates outputs.** This relates to → intelligibility.

Depending on the nature and use of the AI system (AI tool used in research versus tool used in healthcare), descriptions may need to fulfil legal requirements. In the EU, the AI Act ([EU, 2024a](#)) stipulates requirements for “technical documentation” (Article 11 and Annex IV) for AI high-risk systems. In addition, specific documentation requirements may need to be addressed under other applicable law (e.g. the EU Medical Devices Regulation; [EU, 2017](#)) for AI-enabled medical devices software.

Below, we outline above **evidence needs for transparent AI systems used in the health domain**. These considerations may support documentation needs as required and facilitate collaboration of various actors and communities across the value chain and along the life cycle (→ AI actors and communities).

## Evidence on algorithm-to-model transition

### **Parent terms:** Transparency – Transparency of AI systems: evidence

All *assumptions* and *decisions* made in the transition from the concept phase to a usable model that successfully recapitulates a relevant real-world problem (→ **algorithm-to-model transition**) should be properly documented and recorded, including the reasoning behind these.

This includes for example data as well as modelling decisions as well background concerning the → **conceptual relevance** and/or → **contextual relevance** of the AI system: scientific and other assumptions that underpin the design and use of the system, including the strength of scientific evidence supporting these.

In case of AI systems used for healthcare (e.g. diagnostics or clinical decision making) this will specifically concern the → **valid clinical association / scientific validity** for medical devices and in vitro diagnostic medical devices.

## Evidence on data

### **Parent terms:** Transparency – Transparency of AI systems: evidence

Evidence on data may include:

- **Data modality, input-output data:** Information on → data modality of input as well as, where applicable, output data (→ **input data**, → **output and output data**).
- **Data provenance:** Clarity concerning → data provenance of development data, i.e. how these data were acquired, collected and which sources were used. Development data typically concern → **training data**, → **validation data** and → **testing data**. Where → **synthetic health data** have been employed, it should be indicated how these were obtained or synthetized and how they were used (e.g. to fill gaps in the data set).
- **Data privacy:** information on how privacy of data was ensured (→ **PRIVACY PROTECTION**).
- **Data quality:** Description of → **data quality** and → **data quality metrics**, including measures taken to ensure data quality
- **Data labelling, features, attributes, proxies:** evidence on how data were labelled (wrong data labels can cause → **bias** and major safety issues), on data → **features**, → **attributes** and → **proxies**.
- **Data processing / wrangling:** evidence on data processing steps and routine approaches or platforms used.

There may be legal requirements concerning data and data governance, e.g. the EU's AI Act Article 10 ([EU, 2024a](#)).

## Evidence about frameworks and processes used

### **Parent terms:** Transparency – Transparency of AI systems: evidence

Organisations and relevant actors involved in the development, deployment and use of AI systems should be transparent to the extent possible about the general frameworks and processes they use (see also → **ethics code & governance / management**).

This includes for instance

- internal → **ethics codes** for summarizing an organisation's approach to ethical and trustworthy AI,
- general governance and management approaches and more detailed and targeted processes, practices and/or standards that are relevant in the context of → AI risk management, → AI safety and → NON-MALEFICENCE.
- Depending on the situation, approaches such as → ethical evaluation of AI, → AI impact assessments (AI-IA) and → fundamental rights and algorithm impact assessments.

## Evidence on intended use, applicability and limitations

### **Parent terms:** Transparency – Transparency of AI systems: evidence

To minimise errors, failures and → foreseeable misuse, it is critical to provide sufficient information about the intended use conditions and limitations of the AI system.

This concerns a description of

- purpose and → intended use of the system and, where applicable,
- → instructions for use, considering also aspects of → use environments (e.g. hospital, home care) and → use context (e.g. clinical workflow, pathway etc.) as well as relevant minimum requirements and constraints concerning the safe and/or effective use of the system.
- Moreover, specific *limitations* (e.g. inherent uncertainties or → intrinsic incompatibilities and 'trade-offs') as well as *non-applicability* (e.g. to patient groups) should be outlined (→ applicability and limitations).

Depending on the AI system, the following points might be relevant as well, in particular where these impact on use or potential use errors:

- describing → **bias**, in particular unavoidable bias (→ intrinsic incompatibilities and 'trade-offs')
- describing *unavoidable risks*, including a rationale for why these are considered *acceptable*, e.g. when weighed against benefits such as efficiency and equity (→ NON-MALEFICENCE; → BENEFICENCE). This concerns in particular healthcare products for which a benefit-risk profile will be elaborated during → **clinical evaluation** and as part of risk management procedures (→ AI risk management). Information on bias and residual risks will help with integrating the system in workflows.

Notably, information on use aspects may need to be revisited due to experience made in real-world → use environments and → use contexts, after deployment of the AI system. Updates may need to involve various actors or communities along the → AI evidence pathway) and may be prompted by:

- Experiences on the impact of AI systems on the patient-physician relationship (→ upholding a **trustful patient-physician relationship**) and, in particular, guardrails to be applied to avoid a deterioration of this relationship.
- New insights into trade-offs concerning → **interpretability** and **explainability** and → **model performance** requirements that emerge in a specific use context. It may also concern experiences with → **intrinsic incompatibilities** and **trade-offs**, in particular where new post-deployment/post-market information during real-world use has been gathered.
- Experiences in regard to → **applicability** and **limitations**, in particular in regard to new → **use contexts**
- Experiences on interoperability requirements (→ **evidence on interoperability** and **value chain elements**) which may change depending on real-world → **use environments**.

**Parent terms:** Transparency – Transparency of AI systems: evidence

### **Interoperability**

AI systems, like other digital solutions, will typically be embedded in specific → use environments with (existing) infrastructure. Depending on use, AI systems may also interact with other non-infrastructure devices (e.g. other medical devices or internet-of-things (IoT) devices). Integration typically is associated with processes, protocols and procedures for maintenance and proper functioning of digital infrastructure, including cybersecurity-related protocols and measures. Thus, for such integration to succeed, sufficiently detailed information on technical **interoperability** requirements of the AI system is needed. Briefly, interoperability should touch on key value chain elements. These are:

- a) value chain enablers: enabling IT infrastructure, enabling technologies and cybersecurity
- b) value chain assets / values: data (format, exchange protocols), models, other AI systems and services or processes associated with the use of the AI system

Technical details will depend on the specific AI system. For an ontology of value chain enablers, see [Reina & Griesinger, 2024](#).

Interoperability however is not only about technologies and their interaction. Interoperability should may also include indications of necessary **interaction** and **collaboration of relevant actors** → AI actors and communities across the → value chain of AI in health and along the → AI life cycle in health.

### **Transparency of value chain elements that have been purchased or otherwise acquired**

It is increasingly common that AI developers use readily available value chain elements such as data packages that have been purchased or otherwise acquired or pretrained models provided by relevant model libraries ([Reina & Griesinger, 2024b](#)) which are then subsequently refined by the developers using transfer learning ([Chassaigne et al., 2024; 2025 - in preparation](#)).

In such cases, these value chain elements should be adequately described, in particular in cases where their properties may impact patient safety (→ NON-MALEFICENCE; → AI safety; → AI risk assessment; → risks).

This includes:

- → data provenance of development data (i.e. the origin of data, when and how they were compiled and composed)
- Data processing steps including whether data gaps were filled with synthetic (health) data
- other → data quality aspects and → data quality metrics
- information on whether data were assessed for potential data biases that might be propagated in models, leading to → algorithmic bias (see also → bias)
- information on → AI technique and → AI typology of (pretrained) → machine learning models (including → foundation models and → generative AI)
- information on whether and to which extent (pretrained) models are interpretable or require additional (e.g. post-hoc) explainability techniques (→ interpretability and explainability).

Such information will require collaboration and information exchange along the → value chain of AI. The efforts will however pay off in terms of a clearer understanding of responsibilities and accountability (→ accountability structures, attribution of (distributed) responsibilities) and will support a proactive and forward-looking approach to → generating and evaluation evidence on benefits (see → BENEFICENCE).

## Information on training requirements

### Parent terms: Transparency – Transparency of AI systems: evidence

Latest at the point of deployment of an AI system, there should be sufficient information on the training needs for persons using or operating the AI system as intended (→intended use) in a safe and secure manner.

Depending on the AI system, its innovativeness and design as well as its purpose, the following points should be considered when determining training required for the safe and secure use of an AI system:

- Purpose: → intended use, → instructions for use, → use context, → use environment, → foreseeable misuse
- → usability considerations and feedback on usability, in particular where usability problems may cause safety issues (→ AI safety; → communication)
- aspects of interoperability and data or cybersecurity (→ evidence on interoperability and value chain elements)
- information on how to conduct → post-deployment monitoring (non-healthcare AI systems) or → post-market surveillance, market surveillance, corrective action (AI systems used in healthcare) of → model performance (including appropriate → performance metrics) and, especially for systems used in healthcare, on → incidents and → adverse events.
- aspects of use that relate to → applicability and limitations (e.g. specific patient groups), residual → bias (including → intrinsic incompatibilities or ‘trade-offs’ (e.g. optimisation for the positive group).

## Intelligibility

### Parent terms: Transparency – Transparency of AI systems: evidence

With intelligibility<sup>34</sup> we refer to the **desideratum** that AI systems and their underlying → machine learning models are made transparent and described in a way that enables various people (not necessarily only experts) to gain sufficient insight into their functioning and understand their input-output relationship, i.e. how and why specific data lead to specific outputs. Intelligibility is essential for *trust*, for high-risk and high-stake applications such as healthcare and other health domains (in particular health research, public health).

Intelligibility refers to a broad umbrella concept of **understanding AI** (see WHO 2021a<sup>35</sup>; Wade & Bansal, 2019). Such human understanding will be primarily based on the desirable property of models to exhibit interpretability (→ interpretability and explainability), i.e. the degree to which a model *affords* an observer to understand the cause of a decision (Miller, 2019; Biran & Cotton, 2017), or how well humans can comprehend the prediction model encapsulated in a → machine learning model (Burkart & Huber, 2021). In case of highly opaque or ‘black box’ models (e.g. based on → deep learning), interpretability may be attainable only by specific *post-hoc* techniques or models that aim to explain specific outcomes and – to an ex-

<sup>34</sup> Notably, some papers have occasionally equated intelligibility with interpretability or understandability, e.g. Lou et al., 2012. We suggest different notions for the two terms.

<sup>35</sup> The WHO guidance states that intelligibility requires transparency and explainability. While we agree in principle with this argument, the approach omits the well-established term of interpretability (presumably subsuming it under the concept of explainability) and also neglects the need for expressing technical interpretability findings or results from explainability models in terms of adequate scientific explanations. We hence add scientific explanations and clearly communicated explanations as requirements of intelligibility.

tent – may shed light on how the system works. These techniques are referred to as ‘explainability’ methods, i.e. methods that enable explanations. Examples are *occlusion sensitivity* (Petsiuk et al., 2018; Valois et al., 2023) *SHAP*, *LIME* (Salih et al., 2024) or *Grad-CAM* (Selvaraju et al., 2016).

The approach of inferring *ex post* which features may have caused specific outputs relates to '**abductive reasoning**' or inferring the "*best possible explanation*" for a phenomenon or a fact (Harmon, 1965; see comment in Salmon, 1989). With abductive reasoning one infers from the fact that a certain hypothesis (e.g. visualised features using Grad-CAM) can explain the evidence (e.g. the predictive output of an AI system) to the actual ‘truth’ of that hypothesis (Harman, 1965; Salmon, 1989). Abductive reasoning is one logical structure of scientific explanations and aligns also with the general notion of explanations being in a certain way ‘model constructs’ binding real-world facts in a meaningful way (e.g. causality) and whose success can be measured by their ability to account for empirical data, predict phenomena (e.g. outcomes), by their generalisability and – preferably – by their simplicity (Wang et al., 2020). These post-hoc approaches of gauging explanations for black-box outcomes have been aptly subsumed under the term explainability with the associated technological program of ‘explainable AI’ (XAI), launched by DARPA in 2017 (for a summary, see Gunning & Aha, 2019; Gunning, 2021).

Key considerations regarding intelligibility are detailed in sections 1 to 4 below.

## 1. Challenges to intelligibility

Depending on → AI technique used, intelligibility is not a given but needs to be addressed before deploying an AI system for routine use: it is a critical desideratum of AI models (see for instance Weld & Bansal, 2019; WHO, 2021a), especially in high-risk application areas such as healthcare: AI systems need to be *intelligible* for a variety of reasons (section 2), including ethical reasons (e.g. need for informed consent, absence of discriminatory bias) and legal reasons (e.g. accountability, contestability, liability), but also to empower developers in regard to improvements of models: if we do not understand a system, we cannot sufficient comprehend its errors and shortcomings, which complicates addressing these in an effective manner.

The challenge to understand AI systems and their outputs stems basically from two factors: their ‘opacity’ and their ‘alieness’:

- **Opacity:** depending on the type of → AI technique used, models offer more or less insights into how they work. Black box models (e.g. → artificial neural networks) are not inherently intelligible and in most cases their input-output relationship is not readily interpretable.
- **Alieness:** the causality of input-output relations in AI systems may be difficult to understand for humans since these do not recapitulate human cognitive processes and is not based on human reasoning. AI system behaviour may be fundamentally ‘alien’ to us (Weld & Bansal, 2019). The fact that AI outcomes may match decisions achieved by humans are no proof that both identical outcomes have been achieved by the same logical process. Thus, assumptions by people concerning possible causes for outputs of an AI system may be deeply misleading and need to be checked carefully. This illustrated by → *shortcut learning*: the system produced ‘correct’ outputs, but these are based on analysing inappropriate information. Thus, also in situations where there is no element of **surprise**<sup>36</sup> (the output after all seems correct), it is important to try to interpret or explain an AI system. While we may struggle to interpret the behaviour of other fellow humans at times, we share the same world-mapping and interpreting ‘device’ (i.e. our brain) which continuously updates an internal representation of the outer macrocosm (Braitenberg & Schüz, 1991), with cognitive approaches being also to some extent ‘hard-wired’ and thus universal to all humans. This facilitates a comprehension of human behaviour. This is not the case for AI, which

---

<sup>36</sup> Surprise is the primary ‘cognitive trigger’ of seeking an explanation; see Reisenzein et al., 2017.

aggravates the incomprehensibility of AI outputs and impacts trust in AI and thus the willingness to adopt and rely on AI systems (→ TRUST AND TRUSTWORTHINESS).

## 2. **Intelligibility as a desideratum: reasons for wanting to understand AI systems and their outputs**

As observed previously (e.g. [Doshi-Velez & Kim, 2017](#); [Burkart & Huber, 2021](#)), there are a number of reasons why one would like to understand the outputs of AI systems. Some are related to technical aspects, e.g. the need to improve a system. Others are related to → TRUST AND TRUSTWORTHINESS of AI systems: the willingness of humans (individuals, communities, organisations) to adopt and routinely *rely* ([Deley & Dubois, 2020](#)) on AI technology (or other technology for that matter) depends on trust, essentially belief and confidence that such systems work reliably and that they not only make the right decisions, but make these decisions *for the right reasons* ([Weld & Bansal, 2019](#)). Such belief needs to be underpinned by sufficient evidence on:

1. **Causality:** A primary motivation for understanding AI systems is **causality** itself: explanatory understanding aims generally at *causality* (apart from ‘procedural’ explanations, e.g. an explanation how to fix a flat tire). In the context of AI, we want to understand specific properties in the → features or → attributes of the → input data that determine an essentially mechanical cascade of algorithmic events (i.e. the ‘functioning’ of the AI system) within the trained → machine learning algorithm that leads ultimately to a specific output (→ output / output data).
2. **Performance and generalisability:** Understanding general causation of an AI system will support confidence in its performance: good performance on its own does not necessarily prove that a system comes up with accurate outputs for the *right reasons*. Interpretable or inherently ‘intelligible’ AI models ([Weld & Bansal, 2019](#)) afford such understanding (→ **interpretability**) without the need for auxiliary (→ **explainability**) techniques or secondary models ([Rudin, 2019](#)). Similarly, although generalisability to unseen data can be tested to some extent, understanding how an AI system generates outputs will lend confidence to and greatly facilitate a grasp of its generalisability.
3. **Ethical reasons:** an intrinsic property of the human condition is to constantly seek explanations for phenomena. Explanations give a sense of certainty and predictability of possible future events and, importantly, create *trust* in artefacts which we create to aid us. Explanations are a societal need. This is reflected in EU legislations ([GDPR, EU, 2016](#)) which stipulates that everybody has a right to explanation in case automatic decision-making processes are employed for processing → personal data (which includes personal health data). In healthcare, explanations are critical for → ensuring the means for free and informed consent. Adequate explanations that are communicated in an understandable manner to patients are required for the latter to make an informed decision about a medical intervention. This concerns the closely linked ethical principles of → DIGNITY, FREEDOM AND AUTONOMY.
4. **Contestability and liability:** absence of adequate understanding of AI systems undermines a healthcare professional’s capacity for effective → **human oversight**. This might have consequences for their ‘derivative liability’ (see → **liability**). Lack of intelligibility would also have repercussions on understanding sources of failures (→ **correcting problems and failures**), undermining → **failure transparency** and may, depending on situation, affect → **corrigibility**. Finally, lack of an understanding of how an AI system works, undermines the capacity of affected persons to contest decisions (→ **contestability and challenge**) and to seek remedy (→ **remedy and redress**), e.g. in case of inaccurate diagnostic decisions or clinical decisions concerning treatments.
5. **Correcting AI systems and contesting outcomes:** understanding AI systems will allow their improvement in a much more efficient way than if their functioning remains a black box. In the latter case, improvements would approximate ‘trial and error’, while in the former case, targeted

adjustments to a model can be made based on an understanding of causative factors for decisions.

6. **As a proxy for criteria that are difficult to quantify:** An understanding of an AI systems allows its examination in regard to other criteria that are difficult to quantify, e.g. safety (→ AI safety), fairness and non-discrimination and bias (→ FAIRNESS).

### **3. Understanding AI systems and their outputs requires some sort of explanation**

Clearly, by saying that we seek an understanding of AI systems and their outputs we imply some sort of **explanation** of how (specific) inputs are processed by a ‘chain’ of algorithmic events to yield (specific) outputs, i.e. how and why these outputs are produced. Thus, a satisfactory explanation of AI systems and their outputs would be of **causal nature**, which – importantly – does not necessarily imply a *mechanistic explanation* – a subtype of causal explanations. Causal explanations allow to reconstruct and hence understand a) *why* specific outputs are produced (including relevant antecedent conditions or properties of data features, e.g. ‘saliency’) and b) enable to *predict* outcomes for a given input data (this should be possible if we know the rules that govern the input-to-output conversion and the specific antecedent properties that are of particular relevance for these rules). Such predictions would allow **testing an explanation** about an AI system, analogous to the testing of a scientific hypothesis, in view of supporting or ‘confirming’ its likely correctness.

Achieving a satisfactory degree of intelligibility may require technical **interpretations** or **post-hoc explanations** (→ **interpretability** and **explainability**) of how a model may function when producing outputs: in case of so-called black-box models<sup>37</sup> that are not inherently intelligible or interpretable<sup>38</sup> ([Weld & Bansal, 2019](#)), specific approaches are used to elucidate how the model might have produced outputs. This involves often the creation of a second *post-hoc* or *ex post* model (surrogate model) to ‘explain’ the first – or rather a model hypothesis of how the first model may produce specific outputs ([Rudin, 2019](#)). This approach, termed → **explainability**, ‘explainable machine learning’ or ‘explainable AI’ (‘XAI’; see [Gunning & Aha, 2019](#); [Gunning et al., 2021](#)) can be seen as an adjunct methodology to achieve *ex post* interpretability of the model under scrutiny ([Lipton, 2018](#); [Miller, 2019](#)) and, in particular, of the causation of specific outputs.

There is a plethora of different explainability methods, some tailored to specific AI techniques, for example addressing neural or Bayesian networks ([Samek et al., 2019](#); [Lacave et al., 2002](#)) or for extracting rules from support vector machines ([Martens et al., 2007](#)). In any case, explainability approaches appear similar to abductive reasoning ([Lombrozo, 2012](#); [Miller, 2019](#)) or ‘inference to the best possible explanation’ ([Harman, 1965](#); [Salmon, 1989](#)) (see also → **explanations of AI systems and their outcomes**).

Intelligibility addresses not only knowledge of *that* (i.e. expressible in form of an IF – THEN relation, e.g. if there is occurrence of a specific image feature, the model makes a positive prediction), but also knowledge of *why* (i.e. *how*) an AI system produces specific outputs, including specific circumstances that may cause the occurrence of a specific output (causality). Knowledge of ‘*that*’ is descriptive, while knowledge of ‘*why*’ is *explanatory* ([Salmon, 1989](#)).

---

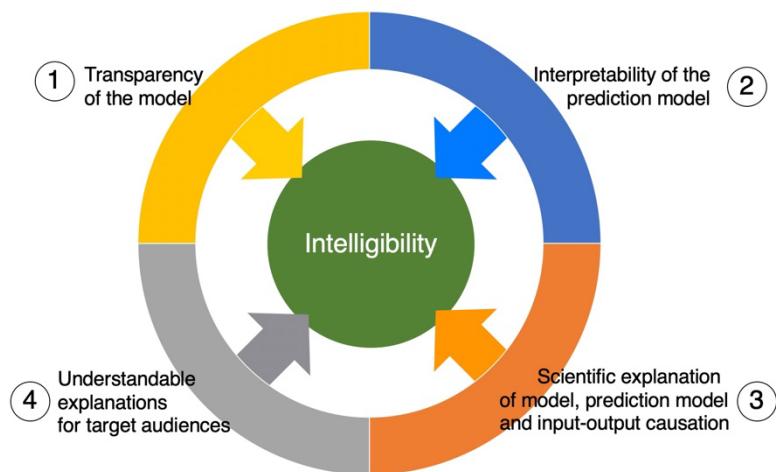
<sup>37</sup> ‘**Black-box**’ models are, as the name indicates, models that inherently are not interpretable. Black-box models are mainly deep learning-based models using artificial neural networks. These may employ thousands of millions of → **parameters** (e.g. weighing factors of inputs between artificial neurons) and are trained on large amounts of data. Thus, their input-output relationship is opaque – as opposed to that of a ‘**white box**’ or ‘**glass-box**’ model, e.g. ‘traditional’ approaches such as a decision trees or linear regression-based machine-learning model (see → **AI technique**).

<sup>38</sup> Discussing ‘simulability’ (i.e. the possibility of a human observer to contemplate an entire model at once), [Lipton \(2017\)](#) argues that there are no ‘intrinsically interpretable’ models: “sufficiently high-dimensional models, unwieldy rule lists, and deep decision trees could all be considered less transparent than comparatively compact neural networks”.

#### 4. Intelligibility entails four elements

We argue that intelligibility requires four elements (**Figure 14**): **(1) transparency** of the AI system and in particular its model; **(2) interpretability** and explainability of the prediction models and input-output causation; **(3) scientific explanations** that provide a coherent logical argument with, ideally, testable predictions (e.g. concerning key characteristics leading to outputs); **(4) understandable explanations** that are tailored to target audiences: there is an obvious enormous difference in regard to details and language (jargon) when explaining a model to a fellow data scientist or to a patient.

**Figure 14.** Four elements required for intelligibility of AI systems and/or machine-learning models: (1) transparent information about the model, including how it, on a general level, transforms data into output or predictions; (2) interpretability of the prediction model affording a more detailed understanding of input-output causality or rules or generalisations used by the model; (3) scientific explanations providing a logical argument of transparency and interpretability elements and allowing predicting outcomes based on specific input data; (4) understandable communication of explanations to various target audiences.



Source: own production

- (1) **Transparency:** An obvious precondition for intelligibility is sufficient transparency of the AI system, in particular the → AI technique employed, the → conceptual relevance (of underlying assumptions, knowledge of relevance for the model), the relevance and quality of the → training data that went into creating the → machine learning model and other decisions that are relevant for creating the model: we cluster relevant aspects under the term → algorithm-to-model transition. In case of inherently intelligible models, a sufficient description of the prediction model may already afford → interpretability, without having to resort to → explainability methods.
- (2) **Interpretability and explainability:** the property of systems to afford insights into input-output relations (→ interpretability and explainability). In cases where inherent interpretability is possible, it is only attainable if the model description is sufficiently transparent. In cases where no inherent interpretability is possible (e.g. inscrutable or black-box models), a degree of intelligibility may be attainable through post-hoc explainability techniques. These entail also secondary models to gain interpretability of input-output relations. Explainability may involve abductive reasoning ('inference to the best explanation'), e.g. based on relevant (statistical) observations and thus relates to a specific approach for gaining scientific explanations (→ explanations of AI systems and their outcomes).

- (3) **Scientific explanations:** the formulation of explanations in form of coherent logical arguments that draw on relevant premises, observations or generalizations, on specific antecedent conditions (where applicable) and insights into model input-output causation (→ **interpretability** and **explainability**). Explanations should lead to testable propositions: to gauge their potential veracity by its capacity to predict outcomes based on input data specifications: explanations should provide a testable view on input-output causation. Given the fluidity of the debates concerning (a) *causation* (see for instance Halpern & Pearl, 2005a, b; Halpern, 2015) and (b) *what characterises a ‘scientific explanation’* (Lombrozo, 2012; Mattioli, 2024), explanations of AI models may follow a variety of logical structures (see → **explanations of AI systems and their outcomes**), including
- classical deductive-nomological explanations** (where the premises are known to be true, allowing to deduce a logically ‘true’ statement), **deductive-statistical** or **inductive-statistical explanations**.
  - abductive reasoning** or ‘*inference to the best explanation*’ (Harman, 1965). Explainability typically applies abductive reasoning to determine the most-likely *potential explanation*<sup>39</sup> for AI outcomes and may use *hypothetico-deductive inference* to support an uncertain premise or uncertain premises, e.g. the causal function of a specific feature (see → **explanations of AI systems and their outcomes**).
- (4) **Understandable explanations:** the effective communication of scientific explanations in view of the target audience’s needs and capacities.

## Interpretability and explainability

**Parent terms:** Transparency – Transparency of AI systems: evidence – Evidence concerning intelligibility

### 1. Interpretability and explainability: issues with their use

To interpret originally means to understand, explain or construe, while explanation means literally to make something plain, i.e. understandable. Thus, there is a fundamental proximity of the concepts of ‘*interpretation*’, ‘*interpretable*’ and ‘*interpretability*’ on the one hand and ‘*explanation*’, ‘*explainable*’ and ‘*explainability*’ on the other hand.

This proximity may have given rise to a considerable confusion concerning the use of the two terms in the literature (Murdoch et al., 2019; Beisbart & Räz, 2021). Both terms are frequently used interchangeably (e.g. Gilpin et al., 2018; or Miller, 2019: “I equate interpretability with explainability”; Cutillo et al., 2020). Explainability is occasionally also used as the broader framing concept (e.g. NIST, 2021a, b; WHO, 2021, which does not mention ‘interpretability’ in its comprehensive guidance on AI in health). **Box 5** shows a detailed example of the near-identical use of the two terms in the literature.

---

<sup>39</sup> Salman (Salman, 1989) argues that abductive reasoning or inference to the best explanation only yields *potential* explanations, but never *actual* explanations, i.e. explanations that are demonstrably true (in the logical sense, not the absolute sense!), which is for instance the case for explanations yielded by deductive-nomological approach

#### **Box 5.** Overlap of meaning between interpretability and explainability

The two citations illustrate the close or near interchangeable understanding (underlined) of interpretability and explainability or XAI in the literature:

Linardatos et al., 2021 describe **interpretability** as “*...interpretability is mostly connected with the intuition behind outputs of a model; with the idea being that the more interpretable a machine learning system is, the easier it is to identify cause-and-effect relationships within the system's inputs and outputs.*”

Ali et al., 2023 describe **explainable AI** (XAI) as follows: “*The process of elucidating or revealing the decision-making mechanisms of models. The user may see how inputs and outputs are mathematically interlinked. It relates to the ability to understand why AI models make their decisions*”.

Likely due to the interchangeable use of the two terms, some guidance documents and standards refer only to explainability, without mentioning interpretability. For instance, neither WHO in its guidance on AI in health (WHO, 2021a) nor ISO in its standard 22989 on ‘*AI concepts and terminology*’ (ISO, 2022) mention or list, respectively, the term *interpretability*. However, ISO/IEC provide, in technical report 29119, separate definitions (ISO/IEC, 2020) for both terms. We argue that subsuming interpretability under explainability is counterintuitive, given the epistemological difference between the two (see section 2).

Interpretability is often used in relation to models that are more immediately understandable, e.g. *intrinsically* or *inherently interpretability* models, contrasting these against ‘black box’ models which require considerable effort in order to understand the probable cause for an output (e.g. Dunn et al., 2021). This notion should however not be misunderstood suggesting that ‘interpretability’ applies *only* to white-box models that can be contemplated at once. Interpretability (see section 3) applies to all types of AI models. Further, there are good arguments to question the quasi-axiomatic assumption that specific techniques (e.g. decision trees) are inherently interpretable or have sufficient simulability (i.e. the ability to contemplate a model at once: Lipton, 2017).

Further, it is often claimed that a higher degree of interpretability associated with more interpretable ‘classical’ models comes at the price of a, comparably, lower (predictive) performance (e.g. Ali et al., 2023 emphasis an “*interpretability-accuracy trade-off*”; London, 2019). While it appears correct, that neural networks are particularly powerful for image recognition/analysis tasks, some authors emphasise that there need not be such trade-off if both aspects are sufficiently considered and analysed during model development (Burkart & Huber, 2021) or that, in particular in health, understandability should not be considered a secondary priority (Rudin, 2019; Vokinger et al., 2021). Moreover, there is evidence that less complex and to an extent more interpretable models are not necessarily inferior to deep learning-based methods for specific non-visual AI tasks such as clinical decision making (Soliman et al., 2023).

These varying notions are not aided by the fact that there, currently, there is no formal consensus understanding of what the term interpretability precisely entails, let alone how it can be mathematically described and objectively measured (Lipton, 2018; Murdoch et al., 2019; Linardatos et al., 2021). As pointed out by many authors, it is comparably straightforward to measure → performance (e.g. through → accuracy). Measuring a qualitative concept relating to human understanding of an automaton’s functioning remains so far more elusive.

## **2. Disentangling interpretability and explainability**

Careful examination of the underlying concepts shows that while both terms are closely related, they relate to different epistemological concepts in relation to gaining an understanding about a model (e.g. “*The terms interpretability and explainability are usually used by researchers interchangeably; however, while these terms are very closely related, some works identify their differences and distinguish these two concepts.*” Linardatos et al. 2020).

- *Interpretability* aims broadly at ***understanding models and their output causation***. Notably, input-output causation should not be equalled to a mechanistic understanding of a model's internal workings). Interpretability is both a desideratum, posited vis-à-vis models as well as the level of understanding afforded by a model to human interpreters. Interpretability appears to focus more "*model explanation*" (Burkart & Huber, 2021, page 253; Murdoch, 2019, page 22073) as compared to explaining concrete instances (specific outputs).
- In contrast, *explainability* entails approaches of constructing ***post-hoc interpretability*** by aiming *explanations* of model outputs in line with the broader concept of *explanation*, i.e. the expounding of facts or data and what circumstances and rules may have caused their coming about. Explainability stretches into the area of scientific explanations and, in particular, ***abductive reasoning ('inference to the best possible explanation')***. Explainability is further often related to ensuring the fairness of model outputs (e.g. absence of discriminatory → **bias**). Explainability can encompass both "*instance explanation*" (Burkart & Huber, 2021, page 253; Murdoch, 2019, page 22073) as well as "*model explanation*" approaches. Thus, explainability methods may, but do not necessarily, elucidate general input-output causation. For instance, so-called 'local explanations' (e.g. saliency maps) do not allow describing the full mapping learning by a neural network (Lipton, 2017).

Taken together, the term 'interpretability' denotes, in particular in the machine-learning community, a broader concept, applying also to black-box models (e.g. Valois et al., 2023: "There's a growing demand for interpreters, tools that decode the influence of input features on a DNN's {deep neural networks} decisions, especially in critical areas like healthcare and autonomous vehicles."). For perspectives on the differences between the terms see for instance Burkart & Huber, 2021; Linardatos et al., 2020; Frasca et al., 2024.

### 3. **Concept description: interpretability and post-hoc interpretability (explainability)**

Against this background we propose the following concept description of interpretability and explainability:

**Interpretability** is the desirable property of an AI system to be accessible to human interpretation and understanding. Interpretability entails two connected perspectives: a) the ability of humans to interpret a model and b) the facility with which a model affords human interpretation. AI models are essentially 'information automata': interpretability aims at elucidating a model's input-output causation (Biran & Cotton, 2017). Interpretability relates to causality, a key element of scientific explanations (→ **scientific explanations of AI systems and outcomes**). Interpretability can be understood as those efforts and methods that focus on the understanding of the model per se (Murdoch et al., 2019). Ideally, interpretability enables predictions about how outcomes will be affected in case input data or computational properties (e.g. hyperparameters) change. Thus, interpretability encompasses key elements of a scientific explanation, including expectability and predictability. Inherent interpretability of AI systems can vary and hence, interpretability encapsulates also the *degree* to which the overall functioning and input-output causation can be understood by humans, i.e. how it produces outputs (recommendations, decisions) why, given a specific input, the system produces a specific output. Notably, interpretability is a desired property of *all models*, irrespective of their complexity or → **AI technique** and architecture, thus encompassing so-called *white box* (e.g. decision trees) as well as *black-box* models (e.g. → **artificial neural networks**).

**Explainability:** Some models are claimed to be inherently more interpretable than others (white box), while others require considerable investigation (e.g. → artificial neural networks). Some models (whatever their makeup: Lipton, 2017) may be so complex as to require specific investigations or techniques, including approaches to elucidate or ‘explain’ *post hoc* (i.e. after the fact) how a model *might have* produced a *specific output* and how the *full mapping of a trained model* looks like (i.e. how it functions in general). Such post-hoc approaches are captured under the term *explainability*. These include auxiliary secondary models that help understanding the function of the primary model (Rudin, 2019). There is an obvious proximity of explainability approaches to the more general concept of (*scientific*) *explanations*<sup>40</sup> of phenomena or facts (i.e. post hoc). However, the term explainability is also used to denote so-called *ante-hoc explainability* of models (see review by Villone & Longo, 2021), capturing efforts to already design models with a view of later understandability. However, given that scientific explanations are, usually about *facts* i.e. phenomena that *happened* (in contrast to ‘explanations’ in the colloquial sense, e.g. ‘I explain to you how to conduct patch-clamp recordings of single neurons’), it might be better to use the term *ante-hoc interpretability* for endeavours that aim at encapsulating, upfront, a certain degree of *intrinsic interpretability* into a model (e.g. through → causal learning or → neurosymbolic AI). We propose that the term ‘explainability’ should be reserved for approaches that aim at constructing an explanatory logical argument or conceptual model of the working of an AI model and, in particular, why and how it produced specific outputs.

Taken together, it is obvious that

- **interpretability does not necessarily entail explainability** (i.e. post-hoc explanations of outcomes; Linardatos et al., 2021). In fact, Gilpin, 2018 suggested that interpretability without explainability may be insufficient.
- Similarly, **while explainability may support interpretability, the results of explainability approaches do not necessarily enhance the understanding of input-output causation** (interpretability). For instance, as remarked by Lipton (2017), saliency maps provide only ‘local explanations’, highlighting regions in a (e.g. visual) input that, if changed, would most affect the output (Simonyan et al., 2013; Wang et al., 2015).

#### 4. The significance of interpretability and explainability

As outlined under → intelligibility, there are profound motivations for a desire to understand how AI models work and how and why they produce specific outputs, i.e. their interpretability and explainability. At the core is the question whether outputs have been produced *for the right reasons*. This entails discovering potential → short-cut learning but also (discriminatory) biases or outdated or false assumptions that may be deeply engrained in the data (→ conceptual relevance). Having assurance that a model produces outputs ‘for the right reasons’ will create trust in the AI system, supporting acceptance by healthcare professionals, patients, health systems, researchers and society. Informed consent is a cornerstone of ethical healthcare: without a basic understanding of the reasons behind an AI recommendation, a patient is *de facto* not in a position to provide *informed* agreement to clinical decisions that involve AI recommendations (→ augmentation).

---

<sup>40</sup> Lipton considers explanations of model behaviour post-hoc interpretability (Lipton, 2017).

A detailed discussion of post-hoc interpretability or explainability methods is beyond the scope of this ontology. In the following we point to relevant reviews and articles, in particular relating to explainability in health and healthcare.

### **Foundations of machine learning interpretability / explainability: philosophy & social sciences**

- [Mattioli and colleagues \(2024\)](#) provide an analysis of explainable AI (XAI) from a philosophical point of view, departing from the notion that a convincing unifying foundation is still missing. The paper provides an overview of the evolution of the concept of scientific explanation and XAI, pointing to similarities and differences.
- [Tim Miller \(2019\)](#) approached the issue of explanations in AI from a social sciences perspective, drawing on concepts from philosophy, psychology and cognitive science.
- [Villone & Longo \(2021\)](#) explore notions of explainability and related concepts (e.g. comprehensibility) and propose a system for organising evaluation approaches for XAI methods.

### **Interpretability – explainability methods**

- For a fast introduction to XAI, [Dallanoce \(2022\)](#) has provided a **brief outline of key methods**.
- [Linardatos and colleagues \(2021\)](#) have provided a **detailed review on machine-learning interpretability methods** (explainable AI)
- [Burkart & Huber \(2021\)](#) survey essential definitions, approaches and **methodologies of explainable supervised machine learning** (which is particularly relevant for AI applications in medical imaging)
- [Holzinger et al. \(2022\)](#) provide a concise **introduction to explainable AI methods**. The article is especially aimed at application engineers and data scientists. The article covers LIME, Anchors, GraphLIME, LRP, DTD, PDA, TCAV, SGNN, SHAP, ASV, Break-Down, Shapley Flow, Textual explanations of visual models, integrated gradients, causal models, meaningful perturbations and X-NeSyL.

### **Interpretability and explainability in medicine**

- [Van der Velden and colleagues \(2022\)](#) provide an overview of explainable AI (XAI) in deep learning-based medical **image analysis**. An overview of XAI methods aimed at clinical practitioners has been provided by [Borys et al., \(2024\)](#).
- Holzinger and colleagues explore explainable AI in medicine in the context of → **causability and multi-modal causability** ([Holzinger et al., 2019; Holzinger, 2021; Plass et al., 2023](#)).
- [Biswas \(2024\)](#) provides a review explainable AI for **disease diagnosis**. Zhang and colleagues (2022) explore applications of XAI in **diagnosis and surgery**.

### **Critical view on the demand for explainability in healthcare**

- Critical views on the need for explainability and associated demands are provided by [McCoy et al., \(2022\)](#) and [Ghassemi et al. \(2021\)](#).

## Explicability

**Parent terms:** Transparency – Transparency of AI systems: evidence – Evidence concerning intelligibility

Explicability has at least two different notions. This may obviously create confusion. We explore below these notions and suggest a way forward concerning use of the term:

### **Notion 1: explicability as a composite ethical principle**

Explicability has been used with a very specific meaning. Floridi et al., (2018) have proposed the term explicability as a novel *ethical principle*, composed of notions of → *intelligibility* and accountability (see also Floridi & Cowls, 2019). Explicability in this sense has been used as one of the four ethical principles outlined in the ethics guidelines of the European Commission's high-level expert group (EU HLEG, 2019). It should be noted that the concepts of intelligibility and accountability are, in the literature, typically discussed in association with separate ethical principles, i.e. intelligibility in the context of *transparency* and accountability typically in the context of *responsibility* (Jobin et al., 2019; Ryan & Stahl, 2021; see reference section of → TRUST & TRUSTWORTHINESS). Notably, there is no consensus on whether explicability should be regarded as an ethical principle (Robbins, 2019; Herzog, 2022). The possible problem of confusing *explicability* with → *interpretability* and *explainability* has been noted by Herzog, 2022. We argue that explicability as a composite concept is useful **in particular in critical contexts and situations**. So-called "explicability measures" (e.g. → *traceability*, → *auditability and auditing*) and transparent → *communication* on system capabilities have been suggested as practical means to confront situations where → *interpretability* and *explainability* of the prediction model cannot be achieved (European Commission high level expert group, 2019). In addition, Adams has argued that *explicability* may be particularly useful in the context of medical decision-making (Adams, 2023).

### **Notion 2: explicability as a synonym of explainability and interpretability**

Explicability is a term that has been used by various organisations and industry (e.g. Microsoft, Google, World Economic Forum), referring to the need that AI should be 'explicable', meaning that it should be comprehensible how and why an AI system provides a given outcomes. Explicability is also used to refer to the question of how AI would act *instead* of human actors. Such usage overlaps with the concepts of → *interpretability* and *explainability* and the broader concept of scientific (mechanistic and causal) explanations, including predictability afforded by sufficiently clear and truthful explanations. For instance, Leslie (2019) states that explicability "...literally means the ability to make explicit the meaning of the algorithmic model's result" and associates explicability with concepts like "*interpretable AI: content clarification, understandable explanation, socially meaningful outcome*". Edwards & Veale (2017) use it in the context of obligations created by data protection laws to provide relevant information or 'explanations' to data subjects in situations of data processing-based automated decision making (ADM), e.g. EU's GDPR (EU, 2016) Articles 13 (2) (f) and 14 (2) (g). The confusion regarding usage is not surprising, given the fact that the words 'explainable' and 'explicable' are synonyms in general language. Nevertheless, the current situation is unsatisfactory and would benefit from conscientious usage of the correct terms under the correct circumstances.

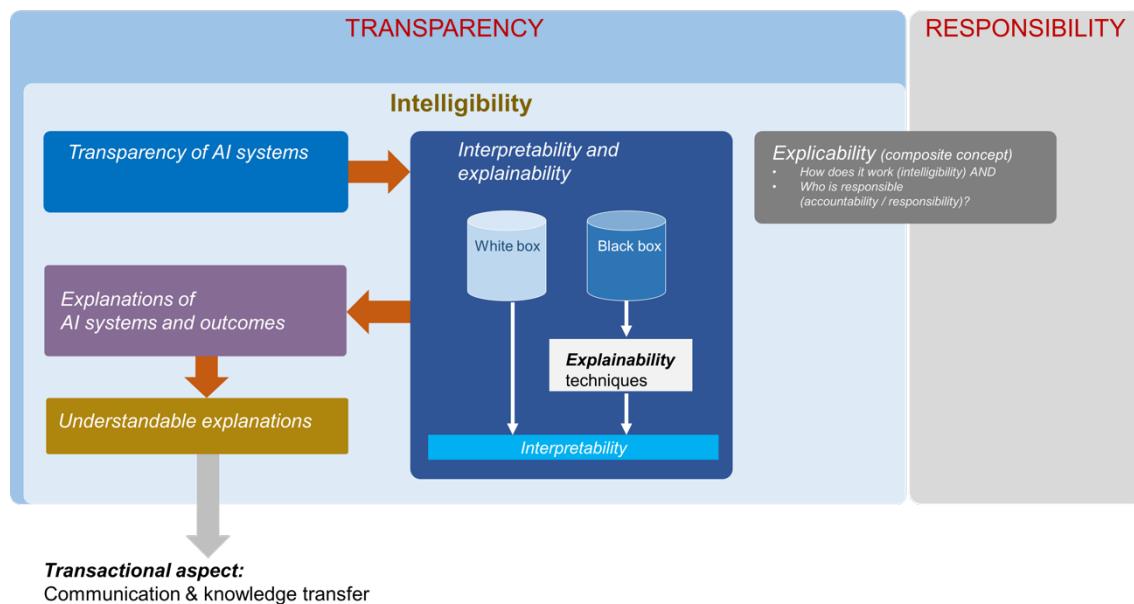
### **Conclusion concerning 'explicability'**

Taken together we suggest

- To use the term '*explicability*' exclusively as denoting the composite ethical principle that combines aspects of *intelligibility* of an AI system and *accountability* (Floridi et al., 2018; EU HLEG, 2019; Adams, 2023). Explicability may be useful in healthcare to address distributed responsibilities for factors that impact on the input-output relationship of AI systems.
- To refrain using '*explicability*' to denote concepts of AI model and AI prediction model → *interpretability* and *explainability*, but instead to use these technical terms.

- To use '→ intelligibility' (instead of 'explicability') to denote the umbrella concept and desideratum of that AI systems and their outputs be sufficiently understandable to whoever may be concerned and has a vital stake (e.g. patients, clinicians, regulators, researchers).

**Figure 15.** Elements of intelligibility, part of the principle of transparency). Explicability is a composite concept that relates to intelligibility (and hence transparency) as well as accountability (and hence the principle of responsibility).



Source: own production

**Parent terms:** Transparency – Transparency of AI systems: evidence – Evidence concerning intelligibility

### 1. From interpretability and explainability to explanations

Results from → interpretability and explainability approaches rarely correspond to a formal (scientific) explanation. They remain at a rather technical level (explaining a prediction model: ‘model explanation’) or may be restricted to specific instances (‘instance explanation’; [Burkart & Huber, 2021](#)).

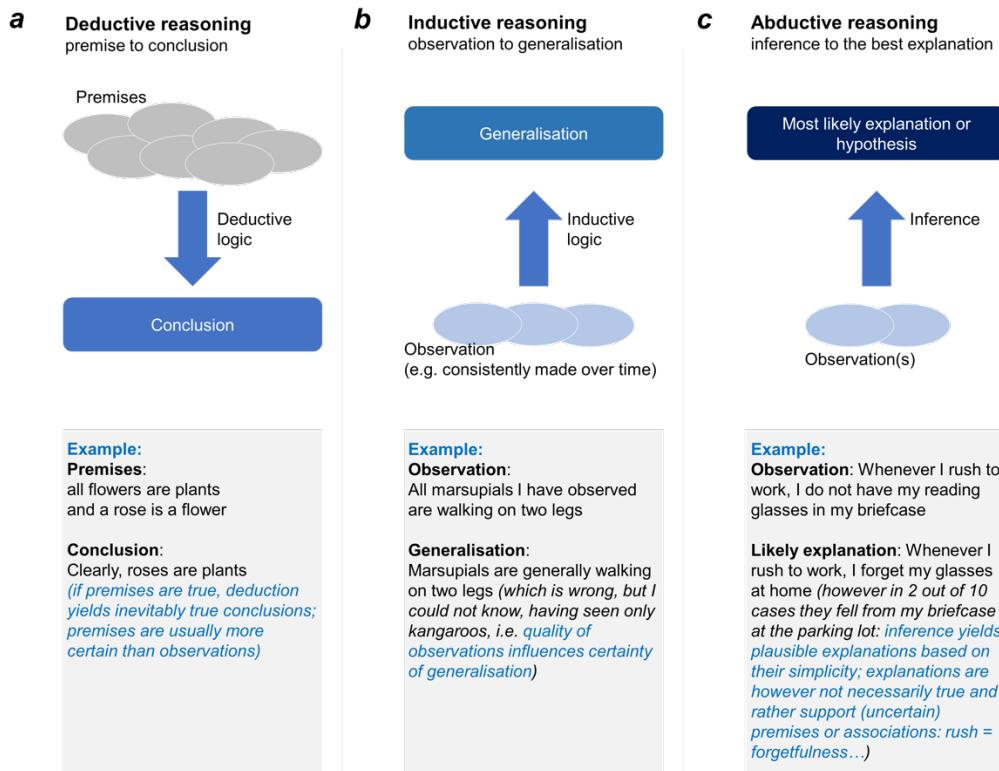
In contrast, (scientific) explanations are related to more complete mental models and concepts about facts, situations or complex systems as well as theory building and. There are two important properties of scientific explanations

1. Ideally a scientific explanation should, where appropriate, **address causality**. It should allow to predict events or phenomena. In the context of AI that would mean to predict outcomes based on specific circumstances (e.g. input data), hence lending support to the likely veracity of the explanation and making its quality as a model of causation testable.
2. Explanations have a **social and transactional quality of communication, clarification and knowledge transfer** – both within scientific disciplines as well as across disciplines and society. Since machine learning → interpretability and explainability are supposed to help elucidating how models work vis-à-vis various communities (e.g. clinicians, patients) to achieve intelligibility, consideration should be given to how relevant findings can be translated into explanatory arguments with a communicative function.

Explanations are “ubiquitous and diverse in nature” ([Keil, 2006](#)), covering instructions of how to do things, justifications of specific choices (“he explained why he did not want to join us for a hike”) and development of theories (“fast endocytosis may be a necessity for maintaining high rates of synaptic transmission”). Typically, explanations involve a logical argument laid out in a series of sentences. There have been various views on what entails an explanation; the highly formal Hempel & Oppenheim deductive-nomological (D-N) model has not fared very well in the philosophy of science ([Salmon, 1989](#)): Individual scientists and scientific disciplines rarely arrive at explanations according to the DN model ([Keil, 2006](#)). From a point of view of logical reasoning, explanations may involve deductive, inductive as well as abductive reasoning ([Figure 16](#)).

The study and debate of explanations is constantly evolving (see [Mattioli et al., 2024](#)) and, after the ‘received view’ ([Salmon, 1989](#)) of what constitutes an explanation (i.e. Hempel & Oppenheim), other models of explanation have been suggested since the 1960ies, e.g. inference to the best possible explanation, the statistical relevance model, the pragmatics of explanation, the unificationist view, counterfactual-manipulative explanation or the ‘decomposition and localisation model’ (see [Mattioli et al., 2024; Miller, 2019](#)). What concerns explanations about machine-learning models, abductive reasoning ([Harman, 1965](#)), also referred to as “inference to the best possible explanation” ([Salmon, 1989](#)), may be a useful theoretical framework ([Miller, 2019](#)). While one could argue however that abductive reasoning rather helps forming explanatory hypotheses than providing explanations, current views on scientific explanations emphasize inherent uncertainties and are not any more bound to the rigid formalism of the Hempel-Oppenheim model.

**Figure 16.** Schematic depictions of types of reasoning that may be involved in explanatory arguments. (a) deductive reasoning (see deductive-nomological model), (b) inductive reasoning and (c) abductive reasoning, also referred to as ‘inference to best possible explanation’. Explainability approaches align often with abductive reasoning.



Source: own production

## 2. Scientific explanations: the ‘received view’

Since Aristotle’s classical framework of explanations, the question of explanations has not been in the focus of philosophers for a long time. The debate was reinvigorated in the 20<sup>th</sup> century ([Mattioli et al., 2024](#)).

### Rudolf Carnap’s explication process: degrees of increasing exactitude

What is perhaps not sufficiently considered in reviews on explanations, is that, before Hempel and Oppenheim proposed the DN model, **Rudolf Carnap** laid out a proposal for a pragmatic “**explication process**”, aiming at replacing a somehow unclear or inexact concept  $C$  (the ‘explicandum’ – to be explained) with a more exact concept  $C^*$  (the ‘explicatum’ – the explained one). Carnap presented this process in 1945, but outlined it in more detail only in 1950 ([Carnap, 1950](#)) (**Figure 17 (a)**).

Importantly, Carnap emphasized that the increase in exactness may come in degrees and that it is irrelevant to which part of the language  $C^*$  belongs: how far it would need to involve non-ordinary language was dependent on the specific case at hand. Carnap stipulated several criteria for judging the success of an explicatum  $C^*$ , *inter alia*: similarity to the explicandum, enhanced exactness, fruitfulness (further insights, consistency etc.), simplicity.

This gradual, flexible and non-formalistic approach seems rather appealing today, given the increasing realisation that explanations are not only a ‘product’ but also a ‘process’ ([Lombrozo, 2012](#)).

### The Hempel-Oppenheim deontic nomological model

A more elaborated and perhaps more rigid model of scientific explanations was proposed by **Hempel & Oppenheim** in their seminal essay “studies in the logic of explanation” ([Hempel & Oppenheim, 1948](#); an introduction to the DN model can be found in [Salmon, 1989](#) and [SEP, 2021 – entry: scientific explanation](#)). Hempel & Oppenheim proposed the so-called *deontic-nomological (D-N) model*. ([Figure 17 \(b\)](#)). Other subtypes of explanatory models such as the inductive-statistical (I-S) or the deductive-statistical (D-S) have been proposed by Hempel and Oppenheim (see [Salmon, 1989](#)) ([Figure 17 \(c\)](#)). Explanations of that pattern should not be confused with either *hypothetico-deductive confirmation of hypotheses* or *hypothetico-deductive inference*, where an explanatory argument of assumed truth is used to infer to the credibility of a set of uncertain premises ([Figure 18](#)). Neither of the two latter are actual *explanations*, although both may play a role in → *interpretability* and *explainability*.

The main Hempel-Oppenheim type of explanation is the D-N model. It consists of an ‘**explanans**’ (nomological component; nomos= law) that logically implies an ‘**explanandum**’, a conclusion:

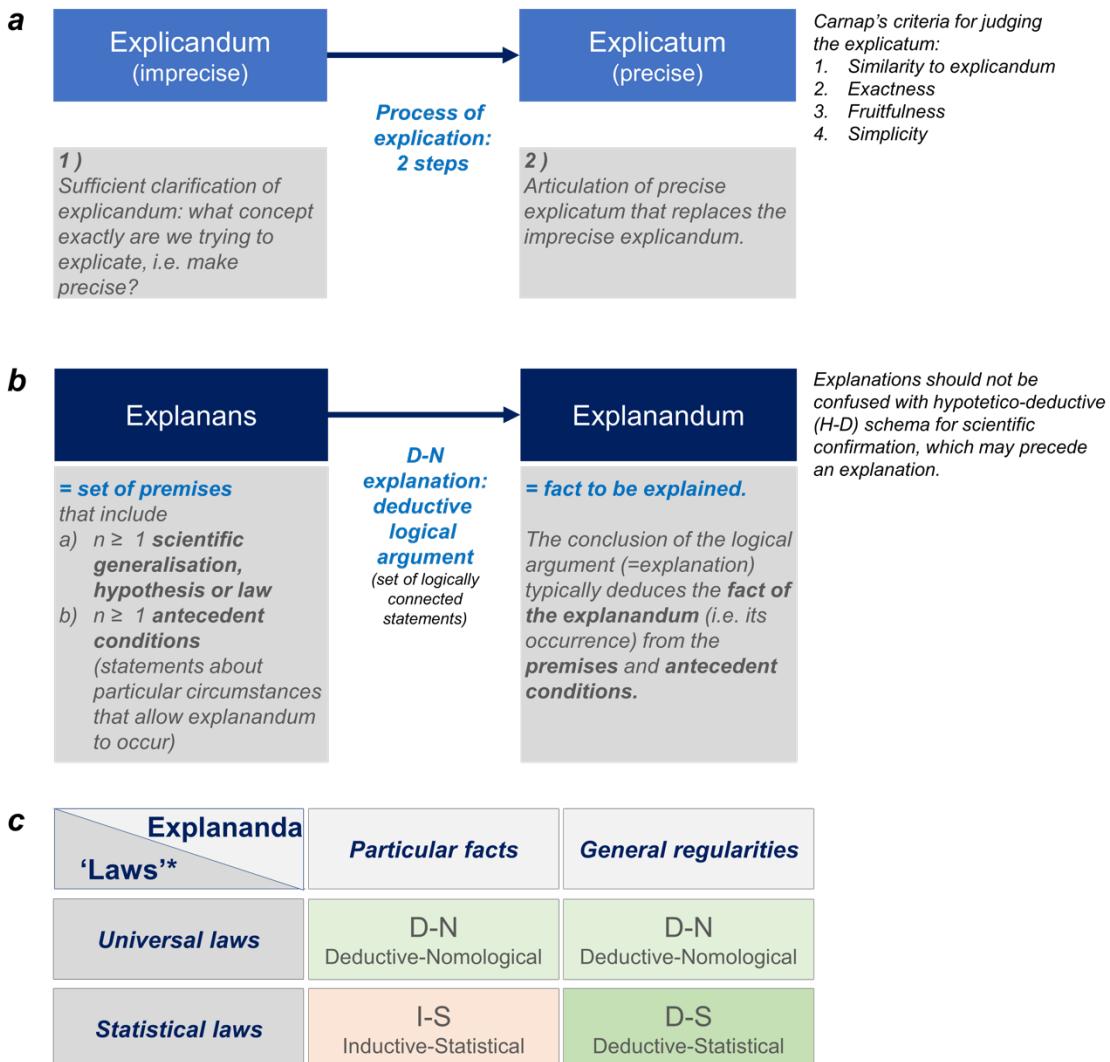
- The nomological component or ‘**explanans**’ consists of a set of
  - i) premises, laws or generalisations  $L$  as well as
  - ii) antecedent conditions or specific circumstances,  $A$ .
- The deductive consequence of this set of premises are then expounded in a logical argument or proof (deductive part), which is the conclusion or explanandum,  $E$ .

[Sadegh-Zadeh \(2011\)](#) provides an **example of a D-N explanation in diagnostic medicine**:

"L      if one of the main coronary arteries of the heart of a human being occludes at time  $t_1$   
       then she suffers myocardial infarction after a short time  
 A1     Hilary is a human being  
 A2     A main coronary artery of Hilary's heart occluded at time  $t_1$  (e.g. then minutes ago)  
 A3     time  $t_2$  is shortly after time  $t_1$   
 E      Hilary suffers myocardial infarction at time  $t_2$ ."

Note that this example provides a causal premise (if occlusion, then infarction), but not a mechanistic one, i.e. there is no mechanistic explanation how occlusion and myocardial infarction are connected through a chain of specific events. Importantly, causal explanations should not be mistaken for mechanistic ones. In medicine, mechanistic explanations may not be always available ([Herzog, 2022](#); [London, 2019](#)). Thus, when discussing explanations about AI systems and models, this constraint should be kept in mind. Clinicians and patients may be satisfied with causal evidence that is associated with uncertainties or even with correlative evidence ([Hagendorff, 2021](#)).

**Figure 17.** (a) Carnap's explication process, (b) Hempel & Oppenheim's deductive-nomological model of an explanation; (c) logical explanation models for universal or statistical laws versus particular or general regularities (adapted from Salmon, 1989).



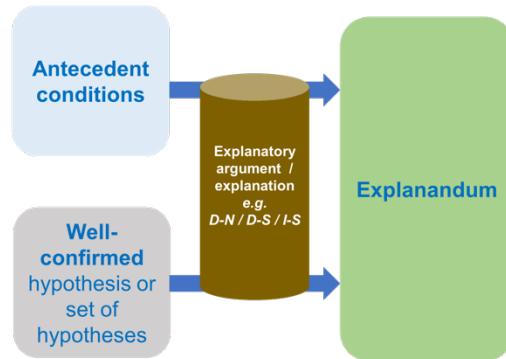
\*) scientific law or established generalisations or hypotheses.

Source: own production

**Figure 18.** (a) Scientific explanation according to the ‘received view’ proposed by Hempel & Oppenheim, using a composite of antecedent conditions as well as laws / generalisation or well-confirmed hypotheses to deduce or infer an explanation. (b) Schematic depiction of hypothetico-deductive confirmation of hypotheses, which should not be confused with an explanation. (c) Hypothetico-deductive inference lens support to uncertain premises, based on an explanation found to be robust. Such inference is not an explanation.

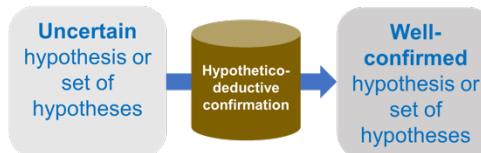
**a    Scientific explanation according to the D-N, D-S, I-S models**

Explanation of a phenomenon that has factual status (=occurred)



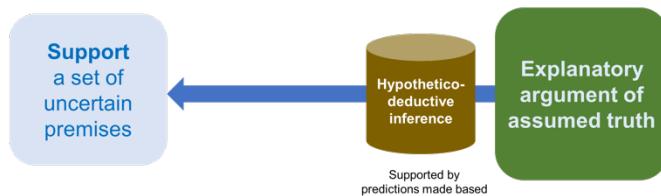
**b    Hypothetico-deductive confirmation (not an explanation)**

Confirm likely veracity of hypothesis



**c    Hypothetico-deductive inference (not an explanation)**

infer from an explanation on the likely truthfulness of premises used for that explanation (via abductive reasoning)



Source: own production

### 3. Explanations of AI systems and their outputs: three fundamental questions

When discussing the explanation of AI systems and their outputs, we propose to discern at least **three fundamental questions**:

- **Explanations of general input-output causation:** Explanations of how the trained → machine learning model *in general* processes inputs to reach an output (recommendation, decision), i.e. what are the ‘rules’ or types of reproducible causal events (even if not identical between chains of input-output connections) that lead from a data input  $D_x$  to an output  $O_x$ . For traditional machine learning model, this ‘functioning’ may be inherently understandable. For black-box models based on → deep learning, there is typically no inherent interpretability, although novel approaches such as → causal machine learning or → neurosymbolic AI may change this.

- **Explanations of specific input-output causation:** Knowledge of general causation should typically afford a comprehension (i.e. explanation) of *specific causation*. With this we mean the understanding of *why* a *specific* output occurred, i.e. how the system processed specific input  $D_s$  data to generate a specific output  $O_s$ . This can also be formulated as being able to conclude or deduce that a specific output *occurred*, because of an applicable set of credible generalisations and *specific* antecedent conditions. This comes close to a deductive-nomological explanation.
- **Counterfactual or contrastive explanations:** This refers to an understanding *why* a generated specific output  $D_s$  was produced *as opposed to alternative one(s)*  $D_s'$  (e.g. why the system produced a positive diagnosis, as opposed to a negative one). This is referred to as *counterfactual reasoning* ( $D_s'$  is then called the ‘*counterfactual case*’; [Lombrozo, 2012](#)) or ‘*contrastive explanation*’ ([Miller, 2019](#)). It involves also an understanding of what alternative rules or generalisations and/or antecedent conditions would have produced  $D_s$ . Formally (i.e. in philosophical logic), counterfactuals are “*propositions or sentences, expressed by or equivalent to subjunctive conditionals of the form ‘if it were the case that A, then it would be the case that B,’ or ‘if it had been the case that A, then it would have been the case that B’; A is called the antecedent, and B the consequent*” ([Hájek, 2001](#)).

Importantly, explanations of AI systems at times use secondary models to ‘explain’ the functioning of the primary model (i.e. → **explainability**). This can be likened to the process of **abduction or inference**, i.e. **the inference to the best explanation**, an approach which however has been considered not a true explanation ([Salmon, 1989](#); see also Rudin’s comment on explainability approaches and their relationship to explanations: [Rudin, 2019](#)). In any case, this illustrates the broad range of logical and argumentative approaches that are employed in the search for an understanding of computational models.

While causal explanations provide an understanding of relationships between **antecedent A** (data) and **event E** (output), they do need not be necessarily ‘mechanistic’. Mechanistic explanations are a specific form of causal explanations, where the emphasis is “*on information about the component parts of a system, their activities, and the spatial and temporal constraints on their organisation in virtue of which they together produce the system’s behaviour*” ([Craver, 2025](#)). Deriving fully mechanistic explanations of neural networks would probably imply an understanding of how each artificial neuron contributed to an outcome at a given point in time. This is clearly difficult or impossible to attain and likely meaningless in terms of advancing explanatory arguments. We argue that such mechanistic understandings are not necessary for a satisfactory causal explanation. As observed by London ([2019](#)), explanations in medicine are not necessarily fully mechanistic either. Useful explanations utilising robust and tested generalisations, rules or ‘laws’ and (statistical) observations of salient antecedent conditions (e.g. data → features; → attributes) will allow predict outcomes for given input data. Such predictions can be used to support or confirm the explanation or the explanatory hypothesis<sup>41</sup>, e.g. through **hypothetico-deductive scientific confirmation** ([Salmon, 1989](#)).

---

<sup>41</sup> “*The degree of success of a mathematical model can be measured by its ability to account for an increasing amount of empirical data, its simplicity and generalizability for novel testable predictions*” (Wang et al., 2020: Computational neuroscience. A frontier of the 21<sup>st</sup> century.)

## Understandable explanations

**Parent terms:** Transparency – Transparency of AI systems: evidence – Evidence concerning intelligibility

Understandable explanations: with this we mean the effective communication of scientific explanations in view of the target audience's needs and capacities, i.e. the translation of knowledge into communicative acts with a certain appeal function that take knowledge of the target audience into account. What concerns an understandable explanation for a data scientist, may challenge a clinician, and what is understandable for a clinician may be challenging for the average patient who typically has neither a background in medical aspects nor AI. It is critical that the scientific explanations are sufficiently simplified so as to enable clinicians to comprehend the decision of an AI system (and possibly veto it ( $\rightarrow$  ensuring human agency and oversight) and to empower patients to make informed decisions ( $\rightarrow$  ensuring the means for free and informed consent). Understandable explanations are important in view of the transactional aspect of explanations as means of communication and knowledge transfer.

## A.6 Responsibility

### Concept description

#### **Responsibility**

With the principle of ‘responsibility’ we refer to the fundamental *ethical* obligation of actors and communities to behave and act *responsibly* when designing, developing, deploying, using and decommissioning AI systems in health. Professional integrity, scientific, technical and clinical excellence and maintaining a quality culture (including risk management) are key to responsibility.

As the word responsible indicates, acting responsibly means being willing and able to adequately *respond* to justified questions and inquiries from stakeholders (e.g. users, patients), regulators or the public. This involves

- working in full respect of the **applicable law**, e.g. in the EU AI Act ([EU, 2024a](#)), GDPR ([EU, 2016b](#)), MDR ([EU, 2017a](#)), IVDR ([EU, 2017b](#)); see explanatory note 1), in alignment with **relevant treaties and conventions** (e.g. Treaty on European Union, 2016, e.g. Article 2; see explanatory note 2; Council of Europe, 2024), in agreement with **regulatory guidance** (e.g. the EU's MDCG guidance) and giving consideration to available ‘**soft law**’ issued by public bodies or bodies working in the public interest on a no-profit basis. This includes non-regulatory guidance like ethics guidelines (e.g. [EU HLEG, 2019](#); WHO, 2021a, b; 2023, 2024) which are typically broader in scope than the normative provisions of specific laws (for more details, see → TRUST AND TRUSTWORTHINESS).
- being able to justify relevant decisions and choices as well as the functioning of AI systems based on robust **evidence**,
- being available to appropriate **scrutiny** (e.g. peer-review or audits; → auditability and auditing)
- **correcting problems and issues** and be transparent about these through effective communication and adequate disclosure
- being responsive to **problems including harms** that AI systems may have caused to persons, property or the environment. This relates to **accountability** and requires embracing the concepts of *contestability* and *challenge*. Importantly, it includes legal responsibilities under applicable liability and negligence law. Thus, like → FAIRNESS, responsibility has a strong notion of *legal* justice.
- ensuring that **everything is done to avoid harm to persons** (see → NON-MALEFICENCE) and to reduce → risks to the extent possible (without negatively affecting benefits, especially in case of AI systems used in healthcare). This includes appropriate → human agency and → human oversight in the interest of ethical values and → AI safety.

Given the complexity of the → value chain of AI in health, responsibilities can be highly distributed among various actors, e.g. in situations where the AI system relied on distributed ‘modules’ (e.g. use of purchased data packages, pretrained models, deployment modes). Thus, responsible acting requires a clear **attribution of responsibilities**. This will allow required actions to be taken in a timely manner. It also requires that actions are **traceable**, first of all within an organisation, but also across the value chain in case various actors are involved (**distributed traceability**).

#### **Translational concepts of responsibility**

Based on above, we structure responsibility in the following translational concepts

- Having awareness of and working in compliance with applicable legal and regulatory frameworks

- Acting with ethical and professional integrity (i.e. without ulterior motives) and to work in alignment with relevant frameworks (e.g. → AI governance, → AI management) and to consider wider impacts on society (→ SOLIDARITY) and the environment (→ SUSTAINABILITY).
- Embracing a culture of quality and, related to this, adequate risk management frameworks or processes (→ AI risk management) in line with legal requirements where applicable
- Foster a constructive dialogue with relevant communities in view of scientific and technical excellence, e.g. through peer review and community discourse and/or by being auditable by non-interested third parties
- Being accountable, i.e.
  - to define responsibilities and the attribution of (distributed) responsibilities within an organisation and across the → value chain of AI and along the → life cycle of AI in health
  - to answer for one's actions. This includes the correction of problems and failures and the provision of relevant information to stakeholders in a timely manner.
- Ensuring → human agency and → human oversight to avoid safety issues (→ AI safety) or negative impacts on → DIGNITY, FREEDOM AND AUTONOMY
- Being responsive vis-à-vis demands, communications and calls by affected persons and/or organisations. This includes enabling contestability / challenge, redress / remedy and legal liability.

#### Explanatory note

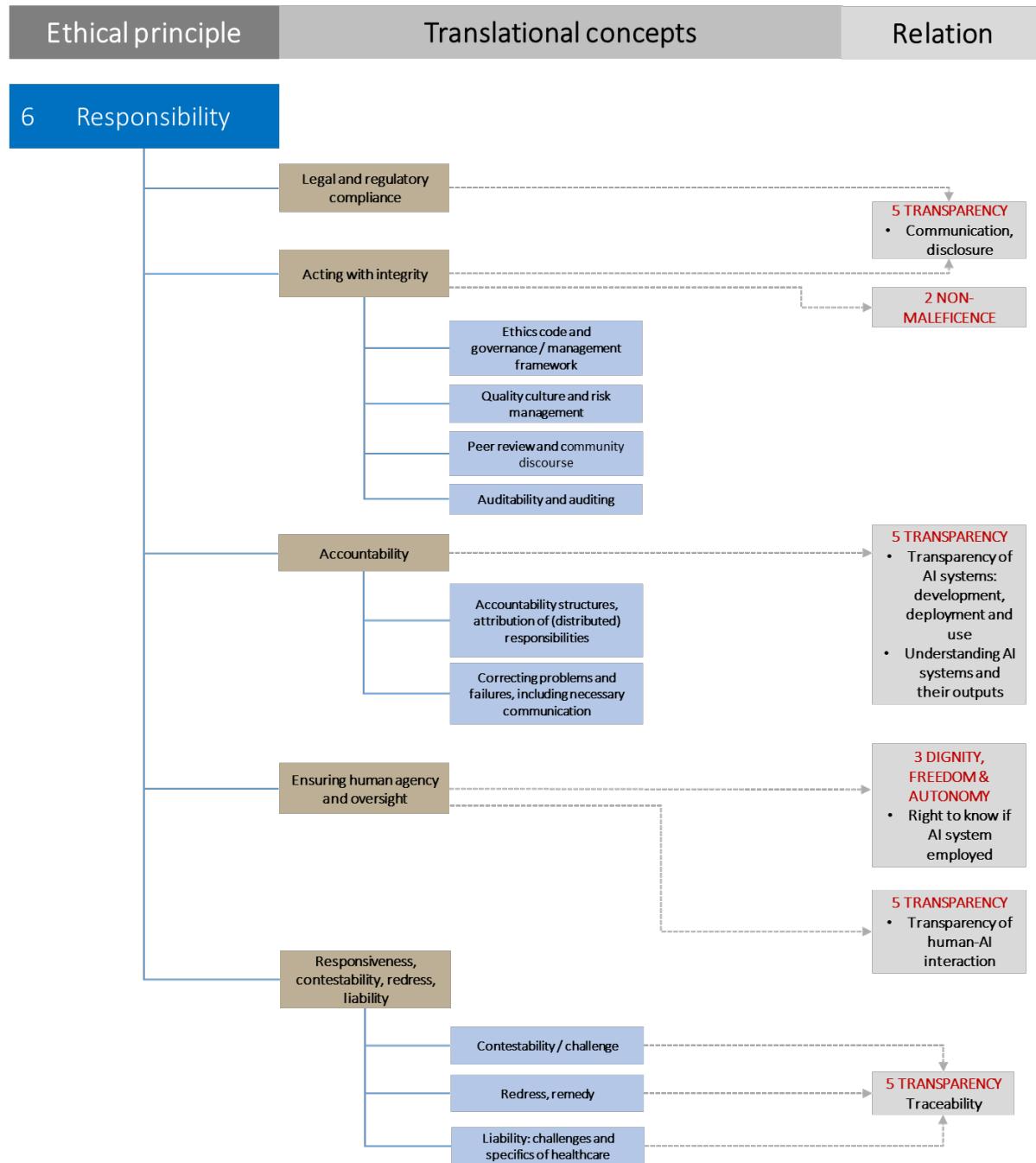
##### **Relevant legislations in the EU**

In a study commissioned by the European Commission, [Lupiáñez-Villanueva et al. \(2022\)](#) outline relevant EU legislations in the area of digital health and artificial intelligence. A summary can be found in Table 24 on page 108 of this document. Note that the document precedes adoption of the EU's AI Act in 2024, which therefore is not mentioned.

##### **Treaty on European Union**

Article 2 of the treaty on European Union (EU, 2016a) states: “*The Union is founded on the values of respect for human dignity, freedom, democracy, equality, the rule of law and respect for human rights, including the rights of persons belonging to minorities. These values are common to the Member States in a society in which pluralism, non-discrimination, tolerance, justice, solidarity and equality between women and men prevail*”.

Ontological organisation of the ethical principle of “responsibility” and its translational concepts. Relations to other ethical concepts are indicated.



## Responsibility: translational concepts

### Legal and regulatory compliance

**Parent terms:** Responsibility

#### **Legal compliance**

Organisations and relevant → AI actors and communities involved in the development, deployment and use of AI systems must take into consideration all **applicable legal requirements**. A study commissioned by the EU Commission on digital health and artificial intelligence outlines (Table 24, section 3.2.1, page 108) key legal requirements within the EU ([Lupiáñez-Villanueva et al., 2022](#)). This may, include relevant impact assessments, e.g.

- privacy impact assessments and Data protection impact assessments (Article 35 of GDPR; [EU, 2016b](#))

or, depending on AI system application area and risk-level

- fundamental rights impact assessments (Article 27 of EU's AI Act; [EU, 2024a](#))

#### **Regulatory documents and guidance**

Regulatory guidance in relation to the AI Act needs still to emerge. For healthcare AI applications, **regulatory information in the area of medical devices** is important. Relevant documents by the IMDRF, the International Medical Devices Regulators Forum should be consulted (Online: <https://www.imdrf.org/>). The IMDRF objective is to accelerate “*international medical device regulatory convergence to promote an efficient and effective regulatory model for medical devices*”. Relevant documentation is developed by several working groups.

The EU's Medical Device Coordination Group (MDCG) provides a wealth of guidance document and relevant information for the EU jurisdiction and jurisdictions of the EU (Online: [https://health.ec.europa.eu/medical-devices-sector/new-regulations/guidance-mdcg-endorsed-documents-and-other-guidance\\_en](https://health.ec.europa.eu/medical-devices-sector/new-regulations/guidance-mdcg-endorsed-documents-and-other-guidance_en)). Particularly relevant to AI systems in health are the documents of the section *New Technologies* such as the Guidance on cybersecurity for medical devices and the guidance *Qualification and classification of software*. While regulatory guidance is not legally binding (in contrast to law), it should generally be followed. It represents a certain extent of consensus within the regulatory community.

#### **'Soft law'**

In addition, relevant '**soft law**', i.e. international guidance, ethical guidelines on AI etc. should be considered. These have typically a broader scope than normative tests (e.g. legislations). While not being legally binding, they may provide more detailed information on non-regulated aspects (examples are integration of AI systems into clinical workflow or automation complacency of AI in healthcare) and/or provide insights into relevant approaches, techniques and the scientific literature. An overview of guidance documents on AI ethical and value-based principles is shown under → TRUST & TRUSTWORTHINESS. Soft-law evaluations such as → ethical evaluation of AI, → AI impact assessments (AI-IA) and → fundamental rights and algorithm impact assessments ([Simoncini & Longo, 2021](#)) may support processes and prepare for legal / regulatory compliance.

### Acting with integrity

**Parent terms:** Responsibility

#### Ethics code & governance / management frameworks

**Parent terms:** Responsibility– Ethics code & governance / management framework

Organisations and relevant → AI actors and communities involved in the development, deployment and use of AI systems should act according to relevant fundamental values or ethical principles and in agreement with professional integrity (e.g. clinical ethics) as well as scientific excellence. It is advisable to inform stakeholders about relevant codes of ethics or frameworks that are followed.

This includes for instance

- internal → **ethics codes** for summarizing an organisation's approach to ethical and trustworthy AI,
- general governance and management approaches and more detailed and targeted processes, practices and/or standards that are relevant in the context of → **AI governance**, → **AI management**
- relevant guidance and codes of medical or other professional associations.

## Quality culture and risk management

**Parent terms:** Responsibility– Ethics code & governance / management framework

Having a culture of quality is central to professional integrity, irrespective of the area of work. In case of product development, assurance of quality is closely linked with → **risk management**. The establishment of quality assurance processes and a risk management system support and complement each other.

The EU's AI Act requires the establishment of a risk management system (Article 9, [EU, 2024](#)) and a quality management system (Article 17). According to Article 17 the quality management system shall include the risk management system (Article 9). The EU's MDR requires a quality management system (Article 10) and risk management system (Section 3 of Annex I, [EU, 2017](#)).

## Correcting problems and failures, including necessary communication

**Parent terms:** Responsibility– Ethics code & governance / management framework

Providers of AI systems must guarantee that AI systems being put on the market or put into service are in conformity with the relevant regulations (e.g. EU's AI Act; [EU 2024a](#)) and are safe when used under correction conditions of use. However, as for all technology, non-conformities, problems and failures may occur, despite the best efforts during the development phase including verification and validation exercises.

In such cases of non-conformities, corrective actions must be implemented and communicated to stakeholders such as distributors, deployers, authorised representative and importers as well as end users (e.g. EU's AI Act, Article 20; [EU, 2024a](#)).

Effective → **post-deployment monitoring** and **post-market surveillance mechanisms** (→ **post-market surveillance**, **market surveillance**, **corrective action**) are prerequisites for detecting problems and solving these through corrective actions.

Relevant general governance and management approaches and more detailed and targeted processes, practices and/or standards may be implemented in the context of processes related to the → **life cycle of AI** in health, the → **value chain of AI** and, importantly, → **AI safety**.

## Peer review and community discourse

**Parent terms:** Responsibility– Ethics code & governance / management framework

Developing and deploying AI systems in a responsible way also rests on sufficient feedback from scientific, technical, clinical and ethical experts, e.g. through peer-review and/or wider community discourse. The active engagement with experts from diverse professional backgrounds will enhance critical evaluation of design, development, and deployment aspects, e.g. decisions concerning → **data**, → **algorithm-to-model transition**, → **intrinsic incompatibilities** and 'trade-offs', → **usability** etc.

Since healthcare products undergo assessment of conformity with essential principles, peer review is particularly relevant for non-healthcare applications. Peer review can for instance include

- a biomedical scientist or clinical assessing the → **conceptual relevance of assumptions**

- a data scientist conducting a code review of an AI model developed by a machine learning engineer to identify potential biases and errors
- an end user (e.g. researcher using an AI system for detecting patterns in health data) may help conducting ‘dry runs’ to evaluate → **usability** and usefulness for the intended purpose.

## Accountability

**Parent terms:** Responsibility

### Accountability structures, attribution of (distributed) responsibilities

**Parent terms:** Responsibility - Accountability

A fundamental challenge of using AI in medicine and healthcare is the complexity of the life cycle and value chain: due to the involvement of many actors along the pathway from developing an AI system to using it under real-world conditions, many things can go wrong and it is difficult to predict how errors upstream may affect results downstream ([Mittelstadt, 2021](#); [Kiseleva et al., 2022](#)).

While for ‘traditional’ health technologies (think of a heart valve implant) responsibilities were less distributed, AI requires a multi-actor environment and collaboration with highly **distributed responsibilities**. Consider for instance a “wearable” medical device used in home care but under supervision by a healthcare provider or healthcare setting: failures may be due to the design of the system, issues due to the → **use environment** and/or the services provided when supervising the device. Thus, AI systems draw on a complex mix of factors, from data / information over IT technologies, modelling to health services.

As emphasized in the WHO ethics guidelines: “*The use of AI technologies in medicine requires attribution of responsibility within complex systems in which responsibility is distributed among numerous agents*” ([WHO, 2021a](#)). Thus, clear demarcation of **responsibility ‘domains’ and attribution of responsibility** is a key challenge for the community involved in developing, deploying, using, monitoring and → **decommissioning / retirement** an AI system in medicine and health (see also [CERNA Report, 2018](#)).

**Attribution of responsibilities** has obviously an impact on potential **liability issues** (→ **liability**). This topic, amongst others, is also highlighted in the 2020 report “Artificial intelligence in health care: medical, legal and ethical challenges ahead” published by the Parliamentary Assembly of the Council of Europe and drafted by its committee on Social Affairs, Health and Sustainable Development (see explanatory memorandum of said document) ([Council of Europe, 2020a](#)). The reply of the Council of Ministers of 2022 takes up this concern ([Council of Europe, 2022](#)).

[Mittelstadt and co-workers \(2016\)](#) have emphasized the importance of traceability (→ **TRANSPARENCY - traceability**) as an overarching concern (see also Figure 1 in Mittelstadt’s 2021 report for various epistemological, epistemic concerns, with traceability as a general concern permeating all other; [Mittelstadt, 2021](#)). Traceability is essential for the improvement of AI systems based on experience. Traceability will, if properly implemented, allow understanding how and where along the pathway things went wrong and what changes need to be made. Such insights are essential for → **AI safety** and → **NON-MALEFICENCE**.

While traceability in a single organisation may be an issue only of appropriate internal quality culture, establishing traceability *processes* in case of highly distributed responsibilities will pose serious challenges. Assigning responsibilities and ensuring traceability through collaboration of all actors in a multi-actor environment that covers the life cycle and spans across the value chain is hence a top priority for safe AI. Such multi-actor traceability has also consequences on responsiveness (→ **responsiveness, contestability, redress, liability**), in case of failures, incidents / adverse events.

## Auditability and auditing

**Parent terms:** Responsibility - Accountability

Auditability concerns an actor's organisation's readiness to be audited. A key requirement for auditability is the availability of documents, records and logs that allow tracing relevant decisions and steps (→ traceability).

Auditing can include a range of activities with the general goal to assess both *processes and procedures* (e.g. quality culture, → AI risk management) but also *individual products*, e.g. to evaluate an AI system's internal workings, including its → algorithms, → DATA, and design processes.

While this doesn't necessarily require the disclosure of sensitive business models or intellectual property, it does involve making internal reports accessible to external auditors (EU HLEG, 2019). Auditing has been suggested as a necessary precondition to verify correct functioning (Mittelstadt B, 2021). Auditing can contribute to → interpretability and explainability, appropriate explanations (→ intelligibility) that support understandable communications about AI systems. It may support → FAIRNESS (e.g. by detecting previously unknown, hidden → biases). Transparent showing of audit results can enhance → TRUST & TRUSTWORTHINESS of AI technology, particularly in case of high-stakes applications that impact fundamental rights or involve safety-critical functions. For such systems, independent auditing is deemed crucial.

In order to facilitate an AI system's auditability, it is important to

- establish proper mechanisms such as ensuring → traceability and logging of the AI system's processes and outcomes
- setting up environments that allow independent audit of the system, especially for applications affecting fundamental rights

In summary, audits of AI systems may help detecting and addressing issues such as:

- → Bias and → FAIRNESS
- → Data quality and → data quality metrics as well as data integrity
- Model → interpretability and explainability that support → intelligibility
- Appropriateness of documentation on AI systems (→ transparency of AI systems: evidence needs)
- Security and privacy of data (→ DATA PROTECTION)
- Performance degradations and potential root causes, e.g. various drifts / shifts (→ drift / shift in machine learning), in particular → distributional drift / shift.
- Model degradation
- Issues with clarity of → accountability structures, notably attribution of responsibilities
- Gaps regarding legal and/or regulatory compliance
- Systemic risks and unintended consequences
- Lack of testing and → verification and → validation

## Ensuring human agency and oversight

**Parent terms:** Responsibility

**Human agency, human oversight, corrigibility**

→ Human agency and → human oversight concerns the capacity of a human actor to exert control over an AI system, e.g. during development (human agency) and exert control over the functioning and outputs of an AI system in specific situations (human oversight), especially where these concern (consequential) decision-making and thus high-risk AI systems. The EU's AI Act outlines that, for high-risk systems, there should be measures for of human oversight that are "commensurate with the risks, level of autonomy and context of use of the high-risk AI system" (Article 14; EU, 2024a).

While → human agency applies to all stages of the → AI evidence pathway (e.g. for decisions during → algorithm-to-model transition), → human oversight is particularly relevant at the post-deployment stage,

during use of the AI system. Human oversight can be realized in various ways, e.g. human-in-the-loop, human-on-the-loop, human-in-command (see [EU HLEG, ALTAI, 2020](#)). As an example, a means of human oversight concerns ‘stop buttons’ or fast interruption protocols in high-stake and high-risk applications such as robotic surgery. Human agency and oversight are a topic of considerable debate.

Based on considerations of fundamental human rights, universal values and ethical principles rooted in human dignity (→ **DIGNITY, FREEDOM AND AUTONOMY**), there is widespread agreement that AI systems should be developed and used in a manner so as to allow appropriate control or oversight by human actors. For example, according to the “meta-autonomy” (or a “decide-to-delegate”) model presented in Floridi et al. ([2018](#)), *“humans should always retain the power to decide which decisions to take, exercising the freedom to choose where necessary, and ceding it in cases where overriding reasons, such as efficacy, may outweigh the loss of control over decision-making. As anticipated, any delegation should remain overridable in principle (deciding to decide again)”.*

There are however applications and situations where a degree of automation is desired. Fully automated decision-making (ADM) means that there is no active → **human oversight**. Further, self-learning or continuously learning AI systems (→ **continuous and adaptive learning**) may complicate oversight as compared to systems that are ‘frozen’ regarding their capacities once deployed.

#### ***Relationship with corrigibility and tests, evaluations and audits***

Human agency and oversight are related to → **corrigibility**, i.e. the possibility to correct AI outcomes under specific situations. Clearly, such corrections should only be done in justified cases, including emergency situations where there is indication that the outcomes are inappropriate and would cause harm: users should not without reason disregard outputs of a well-designed and demonstrably safe and effective AI system.

Tests, monitoring, audits and assessments can also be considered oversight processes and practices. They may be included in → **AI governance** strategies (Jobin et al. [2019](#)) to “*to define and differentiate roles and responsibilities*” ([NIST, 2023](#)). Both the EC high-level expert group ethics guideline and the associated assessment list for trustworthy AI (ALTAI) ([2019; 2020](#)) provide considerations for → **human agency** and → **human oversight**.

#### ***Human agency and oversight in healthcare and other health applications***

AI systems are intended to enhance → **human agency** (see → **augmentation**). Thus, automated decision-making (ADM) is currently not envisaged or aspired to for a variety of reasons, e.g. in the interest of → **AI safety** and to respect human dignity (→ **DIGNITY, FREEDOM AND AUTONOMY**; → **respecting patient privacy**, → **deskilling**). Already non-automated use of AI may have safety implications (e.g. → **avoiding automation complacency**, → **avoiding automation bias**). ADM would likely exacerbate this.

For other health application, ADM solutions might be used. For instance, under specific circumstances (e.g. serious public health threat), health systems may want to implement ADM for patient flow management (based on → **personal data of patients**) and on-time allocation of resources (i.e. for non-healthcare applications).

#### ***Automated decision-making (ADM) and legal provisions: EU's GDPR***

Data protection laws create obligations for information in case of ADM ([WHO, 2021](#), section 5.3) – irrespective of the technology on which ADM is based. The EU’s GDPR ([EU, 2016](#)) for instance stipulates that data subjects **have the right** “*not to be subject to a decision based solely on automated decision making*” (Article 22), while Article 13(f) foresees that, in case ADM is used, data subjects are provided with “*meaningful information about the logic involved*” in ADM as well as “*the significance and the envisaged consequences of such processing of the data subject*”. This concerns also situation where the data have not been obtained from the data subject (Article 14(g)). Goodman and Flaxman ([2016](#)) have called this the “*right of explanation for each subject (person)*”.

The requirement under the GDPR to provide information on the ADM “logic involved” may require, in case AI systems are used for ADM, → **intelligibility**, → **interpretability** and **explainability** as well as methods for providing → **understandable explanations**. Notably, this specific right in relation to ADM does however **not establish a right for explanations, explainability or interpretability of AI in general**.

## Responsiveness

**Parent terms:** Responsibility

With ‘responsiveness’ we refer to the responsibility of relevant actors to *respond* adequately to stakeholders that are concerned about an AI system’s design, that have been affected (e.g. unjustly treated) or even harmed by an AI systems output. This may include situations where AI is used for automated decision-making. We distinguish three concepts:

- **Contestability and challenge:** there are means for stakeholders to *contest* or *challenge* the output of an AI system and relevant AI actors have procedures in place (e.g. as part of their quality approach; → quality culture and risk management) to deal with such challenges in a timely manner.
- **Remedy and redress:** There are procedures in place that *rectify* or *remedy* outputs that were found to be inappropriate (→ correcting problems and failures, including necessary communication). This may include redress (e.g. through financial compensation, free additional diagnostics tests in clinical contexts). Redress will typically be covered by relevant liability legislation. However, responsible parties may also offer redress outside relevant liability circumstances.
- **Liability:** in case where serious consequences or harm have occurred (e.g. due to malfunctions, problems or failures or due to incorrect decisions by healthcare professionals (including ‘commission by omission’; see also → avoiding automation complacency). Liability law covers such situations.

## Contestability and challenge

**Parent terms:** Responsibility – Responsiveness, contestability, redress, liability

The term *contestability* refers to the ability of stakeholders and affected persons to contest or *challenge* the outcome of an AI system. Contestability requires that stakeholders and, in particular affected persons, have the necessary amount of transparent information about an AI system that affects them ([Council of Europe, 2018](#)). This includes

- **Evidence and information on the AI system and how its model was developed** (→ transparency of AI systems: evidence)
- **Information about past failures and the reasons for these** (→ failure transparency). This enables to detect potential patterns of contestable outcomes, enabling to remedy such outcomes and to afford a timely response to contesting parties.
- **Understandable and easily readable information about how the AI system works and how it produces outputs** (→ intelligibility). Depending on AI system, this may relate to the general functioning of a model and the underlying logic of how it produces outputs (predominantly questions of interpretability; “white box models”) or explainability methods (“XAI”; “black-box models”) that are aimed at identifying ‘nomological’ (i.e. robust rules, correlations, ‘laws’) that can support deductive reasoning about why and how specific outputs were produced (enabling to sketch out a possible *causality chain*). Such information should be ultimately summarised in concise and clear → **understandable explanations** (of the deductive-nomological, DN type) whose probable veracity can be checked through the capacity of explanation to produce correct predictions.

Contestability is particularly important for automated decision-making (ADM) systems. While these are unlikely to be rolled out in clinical care (→ augmentation), these may be increasingly used in the context of patient workflow management, data analytics within clinical information systems (CIS). ADM, may, in particular if incontestable, may undermine trust in health systems with possible grave consequences in regard to public health protection, especially in situations of public health crises. Notably, the EU’s GDPR (EU, 2026b) stipulates a right of “data subjects” to have meaningful information about the logic of ADM, which,

de facto, rests on interpretability and/or explainability-based information (see → ensuring human agency and oversight).

However, contestability is also important in the context of non-automated outputs that support decision makers, e.g. health researchers or clinicians, where AI outputs (e.g. recommendations or predictions) may have considerable impact on consequential decision-making on therapeutic choices or qualification of specific treatments for reimbursement.

Alfrink and colleagues have provided sociotechnical features and practices that contribute to contestable AI and proposed a framework for 'contestability by design' (Alfrink et al., 2022): "*Contestable AI systems are open and responsive to human intervention throughout their lifecycle: not only after an automated decision has been made, but also during its design and development.*"

## Remedy and redress

**Parent terms:** Responsibility – Responsiveness, contestability, redress, liability

A key responsibility of organisations being part of the value chain of providing AI systems (see for instance → actors as defined in the EU's AI Act) and/or using AI concerns the **remedy** of failures and problems that might have occurred due to the use of their AI system (Ryan & Stahl, 2021; WHO, 2021) as well as appropriate **redress** where needed and justified in alignment also with relevant legal provisions of the relevant jurisdiction (see → liability).

Remedy includes reversing inappropriate, incorrect, unjust/unfair outcomes of an AI system. Remedy requires that organisations have appropriate communication lines (→ communication) for receiving complaints and claims by stakeholders and users and that there are processes in place to trace issues (→ traceability) and respond in a timely manner (see also → correcting problems and failures, including necessary communication). It also requires that authorities and affected persons have relevant information that allows meaningful → contestability and challenge. This may include understandable communication on the key aspects of an AI system's functioning, i.e. how it produces outputs (→ interpretability). It may require also a deeper understanding of how and why an AI system produces the specific (contested) output (explainability).

In specific situations, remedy may not be sufficient. This includes circumstances where the use of the AI system led to harmful effects and/or damage of persons, property or other aspects (e.g. the environment) and/or where basic rights were violated (e.g. non-discrimination). Organisations need to be able to provide appropriate and visible measures of **redress** in a timely manner.

There are numerous guidelines and ethics papers by private and public organisations that relate also to redress and remedy (see e.g. Jobin, 2019; Ryan & Stahl, 2021 for references. We mention here only a few. The Toronto Declaration ([Toronto Declaration, 2018](#)) focuses on the impact of AI on human rights and provides "mechanisms for public and private sector accountability and the protection of people from discrimination and promotes equity, diversity and inclusion, while safeguarding equality and effective redress and remedy" ([WHO, 2021](#)). The Council of Europe recommendation on the impacts of algorithms on human rights addresses redress and remedy in conjunction with the use of AI systems ([Council of Europe, 2018](#)) led, in 2024, to the Council of Europe Framework Convention on artificial intelligence and human rights ([Council of Europe, 2024a, b](#)). Article 14 concerns remedies and → contestability and challenge. Other relevant conventions in the context of responsibilities are the Oviedo Convention ([Council of Europe, 1997](#)) and The Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data ([Council of Europe, 1981](#)).

## Liability

**Parent terms:** Responsibility – Responsiveness, contestability, redress, liability

Liability, that is the legal responsibility for something, can be categorised in three broad categories: public liability (civil wrongs), product liability and professional liability. A detailed discussion of liability is out of scope of this ontology. Below we discuss, drawing on literature and EU Commission documents, on how AI may challenge liability law.

## **1      The implications of AI on safety and liabilities are subject to ongoing discussion**

Product liability in general refers to the legal responsibility for any harms caused by a product, including the obligation to compensate for damage. Relevant legislative provisions need to be observed depending on jurisdiction. Product liability typically focuses on *defective* products, i.e. that, for whatever reasons, did not work as intended.

It has been argued that AI challenges this approach ([Selbst, 2019](#)): AI inserts a layer of machine agency between *information* or *data* (e.g. a radiological image) and a *human agent* (e.g. a radiologist advising an oncologist) that makes *consequential decisions*. Due to the complex → value chain of AI, a nexus of inter-dependencies is introduced that is highly distributed and difficult to oversee. This could have consequences on determining who is legally liable if things go wrong.

**Box 6.** Brief summary of EU legislations on defective products, developments concerning liability law and AI, lack of clarity around liability as an obstacle for AI adoption in healthcare

- The **EU has currently two legislations on defective products** (Directives 85/374/EEC on liability for defective products (adopted in 1985) and Directive 1999/34/EC with an extended scope of liability to agricultural and fishery products). The EU is in the process of revising these rules, also in view of new technologies such as AI and digital products: [EU Commission, 2024a,b](#) (website ‘Internal market, industry, entrepreneurship and SMEs’ – liability for defective products); see also [European Parliament, 2022](#); see [European Commission, 2022a](#) and [2022b](#).
- From a **general perspective of liability and AI** see the report of the European Group on Ethics ([European Commission, 2018](#)), the IEEE report on ‘Ethically aligned design’ ([IEEE, 2019](#)), the resolution of the European Parliament on ‘Civil liability regime for artificial intelligence’ ([European Parliament, 2020](#)), the comment on civil liability raised by Ebers and colleagues of the Robotics and AI law society ([Ebers et al., 2021](#)) and the European Commission proposal in 2022 for a Directive on ‘adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive)’ ([European Commission, 2020a; European Commission, 2022b; European Parliament, 2025](#)), which was intended to complement the new EU Product Liability Directive. This proposal was withdrawn by the Commission on 11. February 2025 ([Andrews, 2025](#)).
- With regard to **AI systems in healthcare**, the Parliamentary Assembly of the Council of Europe ([2020](#)) stated in its report on ‘Artificial intelligence in health care: medical, legal and ethical challenges ahead’: “*Further debate is needed on the regulatory requirements for privacy, confidentiality and cyber-safety of sensitive personal health data, informed consent and liability of stakeholders*”.
- Similarly, a **study commissioned by the European Commission** (“Study on health data, digital health and artificial intelligence in healthcare”, [Lupiáñez-Villanueva et al., 2022](#)) identified several obstacles concerning the uptake of AI in healthcare, including “*a lack of clarity on liability of manufacturers, health professionals, patients and citizens in AI system services*”.
- Use of new, emerging AI technologies in healthcare (e.g. → generative AI, → foundation models) may further complicate matters in regard to liability ([Shumway et al., 2024](#)).

## **2      Specific challenges to liability arising from inherent properties of digital technologies**

- Irrespective of whether AI systems are a safety component of a specific product or a product in itself (e.g. EU’s AI Act, Article 6; [EU, 2024a](#)), there is debate amongst academics and lawmakers whether or not AI technology may challenge existing approaches and rules concerning liability.
- According to a report of the European Commission ([EU, 2020: COM/2020/64 final](#)) **key challenges** include (1) connectivity, (2) autonomy, (3) opacity, (4) data dependency, (5) complexity of the products

and systems and their interrelation (e.g. cloud-based storage of medical imaging required for a robotic surgery suite), (6) complex value chains and (7) mental health risks linked for instance to the fact that AI systems challenge the primacy of human actors. The level of autonomy may lead to risks where the possibility for correction or opt-out by human actors is affected or may lead to derivative liability of end users (see below). Opacity-related risks concern risks arising from a possible lack of → **intelligibility** and predictability of an AI system output and, hence, the difficulty to understand *causes of malfunction and damages*.

- The fact that AI systems are typically to be connected to the internet (of things) poses **cybersecurity risks**. Cybersecurity incidents in healthcare settings may cause adverse outcomes in patients. Causal connections between incidents and adverse outcomes on patient health are difficult to establish ([ENISA, 2023](#)) and the evidence in the literature is inconsistent ([Reina & Griesinger, 2024a](#)). Cyber incidents in healthcare settings are recognised as an urgent public health problem (e.g. [European Union, 2024c](#)), despite considerable efforts in the past ([PHCAS, 2025](#) – EU horizon 2020 project ‘PANACEA’). A better characterisation framework of cyber incidents in healthcare settings could support targeted cyber resilience measures, establish correlations with health impacts in patients and aid identifying where things went wrong: “*Member States are strongly encouraged to ...share all cyber incident notifications from hospitals and healthcare providers ... Ideally, this should be accompanied by a meaningful characterisation of various relevant incident dimensions, including known root vulnerabilities and effects on healthcare services and patient adverse events.*” ([European Commission, 2025a](#): Action plan on the cybersecurity of hospitals and healthcare providers).

### **3      Distributed responsibilities as a challenge to liability attribution**

A clear allocation of responsibilities has been possible for ‘traditional’ health products ([Mittelstadt, 2021](#)). As outlined above, AI is perhaps challenging this approach. Reasons include the wider spread of responsibilities due to the inherent complexity of the → life cycle of AI in health and → value chain of AI products. Mittelstadt and colleagues have in 2016 alerted to the shift from clearly allocated technical as well as moral and legal responsibilities towards “**distributed responsibilities**” ([Mittelstadt et al., 2016](#)) and raised this issue more specifically in the context of the impact of AI on the patient-physician relationship ([Mittelstadt, 2021](#)). The EU’s AI Act refers to the EU’s Digital Services Act (Regulation (EU) 2022/2065; [EU, 2024b](#)) what concerns liability of providers of intermediary services (Recital 11, Article 2).

### **4      Liability and healthcare-specific aspects**

- There is inherently an **overlap of liability with other ethical topics**, notably → risks, → AI safety and resilience (see → **NON-MALEFICENCE**). Attribution of liability rests on attributions of responsibilities. The CERNA report on research ethics in machine learning ([CERNA, 2018](#)) outlines a fundamental dichotomy concerning
  - *responsibility of the party producing an AI system* (i.e. developer, designer, and particularly the manufacturer) to build AI systems that are safe and secure (ideally, based on the concept of safety and security-by-design: [European Commission communication to the European Parliament, 2019](#)), and
  - *liability of the users of AI systems*. Liability attribution in case of malfunction, error or any harms should be based on an attribution of responsibilities.
- **Organisational users** have specific responsibilities and may have liability, e.g. the responsibility to ensure cybersecurity of hospital networks and hospital-to-cloud interactions. Vulnerabilities that are based on neglect may lead to damage or harm, including for patients (e.g. impaired patient care, delayed diagnosis or treatment; [Reina & Griesinger, 2024a](#)). It is useful to consider three domains of a larger ecosystem of data, information, connectivity and devices in the context of a hospital organisation:

- (1) ***clinical information systems*** (e.g. clinical data, patient data, testing results, radiological images, local storage, cloud storage etc.),
  - (2) ***administrative information systems*** (e.g. patient workflow management, reimbursement, procurement) and
  - (3) ***connected devices*** that may be within or outside the hospital but still depend on data connectivity with the wider (healthcare) digital ecosystem.
- **Healthcare professionals may have at least two types of liability situations** ([UN, 2020](#)): *direct liability* and *derivative liability*. "Direct liability" implies: the healthcare professional or other user commits active deeds, i.e. he/she uses an AI system, and, as a direct consequence, the patient dies or suffers from serious lasting or non-lasting harm. "Derivative liability" means: the healthcare professional or other user is responsible for a so-called "commission by omission", i.e. the failure to fulfil the duty to avert a harmful event. The duty to act may arise from various circumstances, e.g. a statute, a contract, a personal relationship between healthcare professional and patient, the voluntary assumption of care etc.
- [Selbst \(2019\)](#) has proposed **four ways in which AI challenges existing negligence law**: "*1) un-foreseeability of specific errors that AI will make; 2) capacity limitations when humans interact with AI; 3) introducing AI-specific software vulnerabilities into decisions not previously mediated by software; and 4) distributional concerns based on AI's statistical nature and potential for bias.*"

## A.7 Fairness

### Concept description

#### **A brief survey of fairness and related terms**

In essence, fairness is about treating people consistently equally and rooted in ethical values. Fairness is about justice, i.e. the just behaviour or treatment. No one should be “*treated less favourably than other people are in a comparable situation only because they belong or are perceived to belong to a certain group or category of people.*” ([Council of Europe, 2014](#)).

**Fairness is rooted in the inviolable dignity of every human being** (→ DIGNITY, FREEDOM AND AUTONOMY), which, as a consequence requires equal treatment of all and hence non-discrimination or non-distinction. Fairness can also be seen as a desideratum under a utilitarianism and libertarian approach ([Hansson, 2022](#)): fairness ultimately leads to benefits to the largest possible number of people and equality avoids wasting a society's full potential (e.g. John Stuart Mill; essays '*Utilitarianism*'; see [Berger 1979](#); '*The subjection of women*'; see [FarnamStreet, 2024](#)). These notions are highly relevant for fairness in healthcare.

In terms of international **treaties, conventions and guidance**, there are several documents that address fairness and equality ([see explanatory note](#)). Fairness is closely linked with justice in both the ethical sense (treating people justly) and the legal sense (e.g. not violating relevant non-discrimination laws).

There are a number of concepts that are closely linked to fairness.

- Fairness is a lack of **equality** (not treating all people the same) and lack of **equity** (treating specific people in a different, i.e. discriminatory manner).
- Lack of fairness (and hence inequality or inequity) can be considered as an **injustice**: “*Of all forms of inequity, injustice in health care is the most shocking and inhuman.*” (Martin Luther King, Jr. National Convention of the Medical Committee for Human Rights, Chicago, 1966; cited after [The Joint Commission, 2024](#)).
- **Fairness, non-discrimination, inclusion, diversity** and **plurality** are closely related. While there may be subtle nuances between these concepts depending on context, we consider them largely congruent. AI systems that do discriminate against specific groups, do not consider (include) specific groups and hence do not perform for a plurality and diversity of people are not fair. Fairness includes respecting plurality and diversity ([Zowghi & Bano, 2024](#)) and the multifacetedness of humans with every person having inviolable dignity (→ DIGNITY, FREEDOM AND AUTONOMY). Zou & Schiebinger have reviewed various aspects in development and deployment of statistical AI (i.e. not robotic AI) which need to be considered to ensure AI benefits to diverse populations ([Zou & Schiebinger, 2021](#)).
- Fairness has also a **socioeconomic dimension**, relating to the **just and equal distribution** of both **costs** and **benefits**
- Fairness is also discussed in relation to **justice**, e.g. the right to “*contest and seek effective redress against decisions made by AI systems and by the humans operating them*” ([EU HLEG, 2019](#)). In this ontology, we consider fairness dimensions relating to contestability, redress and legal liability under the principle of → RESPONSIBILITY. Similarly, while the same document refers with fairness also to “freedom of choice”, we discuss this notion under → DIGNITY, FREEDOM AND AUTONOMY, in particular rights that are **specific to health contexts, such as the right to know as well as not to know one's medical condition** and the **importance of free and informed consent**, including its prerequisites of information being

communicated in form that is **understandable** for the target audience (e.g. patients) (→ intelligibility).

- Finally, **fairness is typically due to → bias** and results in **systematic error** or **disparate outcomes** (see section “fairness and bias” below).

### **Fairness and AI use in health and medicine**

When developing and deploying AI for health-related applications, e.g. healthcare, health research, public health or health system management, **fairness is about ensuring inclusion**, i.e. **avoiding discrimination** against specific sensitive → attributes associated with people, such as gender, sex, race, ethnicities, sexual orientation, genetic background, disabilities or age (consider for example the lack of paediatric medical devices; see [European Academy of Paediatrics, 2023](#)). In brief, it is about conceiving, **designing and building AI in an inclusive manner, unless there are specific justifiable reasons to do otherwise**. The result of lack of fairness are “disparate impacts” (a term sometimes used interchangeably with → bias; see below) and lack of → health equality & health equity. The WHO ethics guidelines frame fairness slightly broader, including not only non-discrimination, but also notions of neglect, manipulation, domination or abuse, noting that this concept is also “*sometimes called “justice” or “fairness”*” ([WHO, 2021](#)).

### **Fairness and bias**

The root cause of unfair algorithms or the unfair use and implementation of algorithms is → bias:

- either **bias that got encapsulated in algorithms** (e.g. through **selection of biased data**, data labelling, use of sensitive attributes etc. or biased assumptions, (medical) concepts etc; → algorithm-to-model-transition) or
- **bias regarding the way a system is developed used in practice under real-world conditions**, e.g. in a clinical workflow or already when designing specifications for an AI system. This includes unintentional bias (see comment on bias in medical devices [Naqvi & L'Esperance, 2024](#)), including **cognitive biases** (→ heuristics).

Unfair bias can be understood as **systematic error** that affects specifically vulnerable groups. It follows that not all bias is necessarily unfair. There may be deliberate choices to use “equitable bias” for catering sufficiently to vulnerable groups (see → bias). AI systems based on design targeted to specific groups (e.g. patients with rare diseases) could be considered unfair regarding outcomes when applied to most patients (→ universal versus targeted design). Bias can also be unavoidable, e.g. due to the incompatibility of various notions of fairness in predictive model, which cannot be all satisfied simultaneously (→ Intrinsic incompatibilities or ‘trade-offs’).

As outlined above, fairness is a complex concept and, as noted by Mittelstadt, while there is no single universally accepted definition of (non-)discrimination, there is a long history of jurisprudence discussing types of discrimination, goals of equality and thresholds for distribution of goods across groups ([Mittelstadt, 2021](#)). Considerations of fairness and non-discrimination remain difficult in this complex landscape. Further, sensitive → attributes, especially when hidden in → proxies, may be difficult to detect in aggregated and/or linked data sets.

Finally, as noted by Selbst et al., various proposals of fairness metrics may suffer from a narrow conception of the problem, considering machine learning models, inputs (data) and outputs but “*abstract away any context that surrounds*” the system ([Selbst et al., 2019](#)). This observation is particularly relevant in the context of healthcare with its varieties of → use contexts and → use environments and the multiple ways in which AI systems can be implemented in practice or, end up being used in practice (e.g. “off-label” use). → AI practitioners should consider these aspects when collaborating on fair and non-discriminatory AI.

**Figure 19.** Mind map of the term fairness and related terms



Source: own production

#### Explanatory note

There are several international declarations, treaties and conventions that refer to fairness, non-discrimination and equality.

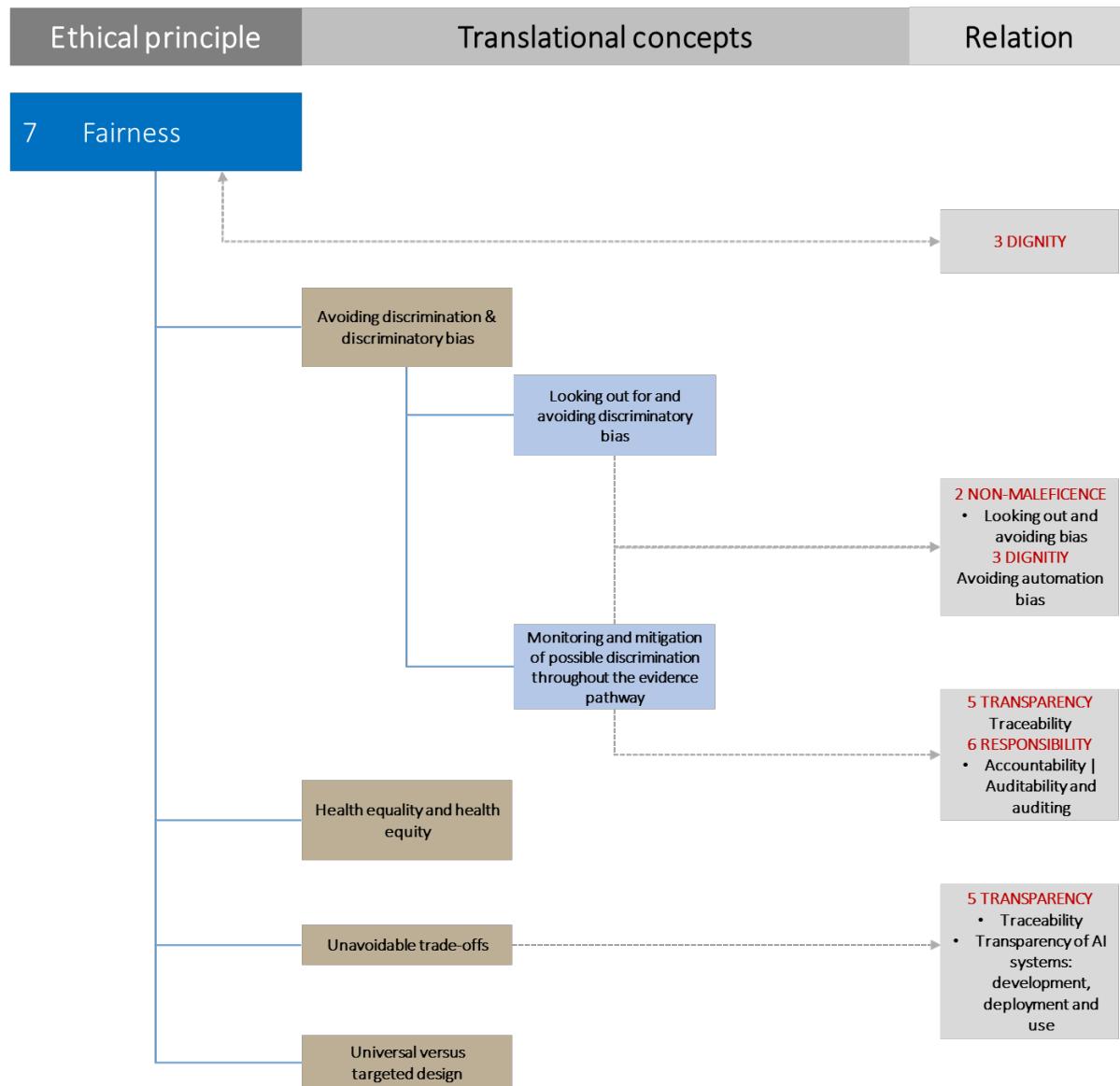
- The **Universal Declaration of Human Rights** ([United Nations, 1948](#)) states that "*All human beings are born free and equal in dignity and rights*" (Article 1); "*Everyone is entitled to all the rights and freedoms set forth in this Declaration, without distinction of any kind*" (Article 2).
- The **European Convention on Human Rights** ([Council of Europe, 1950](#)) lists different sensitive → **attributes** regarding non-discrimination. Article 21 states "*Any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited.*" See also the Charter of fundamental rights of the European Union ([EU, 2000](#)).
- More specific provisions regarding non-discrimination are laid down in the **Oviedo convention for the protection of Human Rights and Dignity of the Human Being regarding the Application of Biology and Medicine** ([Council of Europe, 1997a, b](#)). It states in Article 11 (Non-discrimination): "*Any form of discrimination against a person on grounds of his or her genetic heritage is prohibited*".
- The **WHO in its ethics guidance for AI for health** ([WHO, 2021](#)) understands fairness as a moral demand: "*Ensure that all persons are treated fairly, which includes the requirement to ensure that no person or group is subject to discrimination, neglect, manipulation, domination or abuse (sometimes called "justice" or "fairness").*"
- The **UN Covenant on civil and political rights** ([1966](#)) Article 26 stipulates a need for legal protection from discrimination: "*All persons are equal before the law and are entitled without any discrimination to the equal protection of the law. In this respect, the law shall prohibit any discrimination and guarantee to all persons equal and effective protection against discrimination on any ground such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status.*"

## Term relationship

Related terms:

- Non-discrimination
- Inclusion
- Justice (in both a moral and legal sense)
- **DIGNITY, FREEDOM AND AUTONOMY**
- **RESPONSIBILITY**
- Health equality & health equity
- Bias
- Algorithm-to-model-transition
- Performance metrics
- Attributes
- Proxies
- Intrinsic incompatibilities or 'trade-offs'
- Universal versus targeted design

Ontological organisation of the ethical principle of “fairness” and its translational concepts. Relations to other translational concepts are indicated by grey / stippled arrows.



## Fairness: translational concepts

### Avoiding discrimination and discriminatory bias

#### Parent term: Fairness

Discriminatory bias has been demonstrated on various occasions in relation to AI algorithms used in health (e.g. [Johnson, 2019](#); [Gershgorn, 2018](#)) and is recognised as a major problem of AI-enabled solutions (e.g. [Hoffman, 2021](#); [Rajpurkar et al., 2022](#)). Lack of fairness in AI algorithms can be considered at odds with all four principles of → bioethics: → BENEFICENCE, → NON-MALEFICENCE, justice (→ RESPONSIBILITY) and autonomy (from causation) (→ DIGNITY, FREEDOM AND AUTONOMY) ([Ricci Lara et al., 2022](#)).

We discuss here briefly some key considerations reg. algorithmic fairness and fairness when implementing AI.

#### **Hidden assumptions:**

Some fairness complications stem from a lack of clarity concerning the definition of specific sensitive → attributes, such as “race”. There has been the hidden assumption that race or ethnicity is a reliable → proxy for genetic differences (there is mounting evidence that it is not), which would require attention regarding diagnostic tools and treatments ([Vyas et al., 2020](#)).

Thus, developers should pro-actively watch out for such hidden assumptions that manifest themselves as build-in bias or discrimination that is not evidence-based but either founded in traditional views or → heuristics in model development.

This may relate for instance to diagnostic → algorithms and → clinical practice guidelines used for individualising risk assessments and as a guide for clinical decisions. By encapsulating concepts and → attributes such as race into healthcare decision making, race-based medicine may be involuntarily propagated ([Vyas et al., 2020](#)). Thus, any scientific assumptions or concepts underpinning model development (→ algorithm-to-model-transition) should be transparently documented (→ TRANSPARENCY), so that these can be interrogated and failures can be traced (→ TRANSPARENCY - failure transparency, traceability).

#### **Fairness and underrepresentation of groups in AI development datasets:**

A major issue apart from inherent discriminatory bias in datasets is potential **underrepresentation of specific groups** which leads unreliable and potentially inaccurate outcomes for persons of that specific group (“representational bias”). This may concern for instance gender representation and the global south ([Roche et al., 2021, 2023](#)) or older population groups (so-called “ageism”; [WHO, 2021e](#)) that tend to be underrepresented in data sets used for training AI in health ([WHO, 2022](#)).

The ‘**STANDING together**’ initiative (“standards for data diversity, inclusivity and generalizability”) is an international, consensus-based initiative aiming to work out recommendations for data compilation and data reporting of data used to train AI systems (→ AI-enabled medical device software) used in healthcare ([Ganapathi et al., 2022](#)).

Chin et al., have proposed guiding principles to address the impact of → algorithmic bias on racial and ethnic disparities in health and healthcare ([Chin et al. 2023](#)).

#### **Complexity of fairness when implementing AI in healthcare:**

Apart from hidden assumptions, another difficulty of dealing with fairness in AI may stem from the **inherent multifacetedness of fairness in a complex application domain such as healthcare**. Selbst et al. have described this as the danger to fall in a “formalism trap”: “Failure to account for the full meaning of social concepts such as fairness, which can be procedural, contextual, and contestable, and cannot be resolved through mathematical formalisms” ([Selbst et al., 2019](#)).

Considerations of fairness, inclusion, diversity, plurality are fundamental when **designing, developing, and deploying AI systems** to ensure not only ethical acceptability, but also equitable and **reliable outcomes** ([Zowghi & Bano, 2024](#)) that deliver for the broadest possible number of users and patients, unless

deliberately desired to target a specific group, e.g. to cater specifically for patients with rare diseases/conditions (see → universal versus targeted design) (see Abdallah et al., 2023).

### **Ensuring fairness requires a holistic approach along the entire AI evidence pathway:**

“Fairness by design” requires a **holistic approach along the entire evidence pathway** from conception, design, development, production, deployment to post-deployment monitoring and **requires a collaborative approach of actors of the value chain**. Potential issues and sources of lack of fairness/discrimination should be identified in a forward-looking manner

- at all possible steps of the → algorithm-to-model-transition, e.g. data related bias, fundamental (scientific or clinical) assumptions, modelling approaches and decisions, performance metrics etc. This will allow tackling → algorithmic bias and its impact on equality and equity.
- When considering how to implement the AI system in a given → use environment and → use context, e.g. a clinical workflow or → clinical pathway.

#### **Box 7. Examples of open source toolkits for fairness, non-discrimination and equality**

- IBM is offering **‘AI Fairness 360’** - an open source toolkit to „examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle“. Online: <https://aif360.res.ibm.com/>
- **Fairlearn – improve fairness of AI systems** is “an open-source, community-driven project to help data scientists improve fairness of AI systems.” Online: <https://fairlearn.org/>
- AI & equality provides an **‘AI & Equality Human Rights Toolbox’**, “designed to provide a global community the training, resources, and support to learn from and contribute to a Human Rights-based approach to AI.” Online: <https://aiequalitytoolbox.com/methodology/resources/>

- **Who may be concerned?** AI developers (including data scientists, modellers, life scientists and biomedical scientists advising modellers), AI deployers (e.g. in hospitals and other healthcare settings), clinicians, health information professionals.

### Looking out for and avoiding discriminatory bias

**Parent terms:** Fairness- Avoiding discrimination: inclusion, diversity and plurality

There is ample evidence of discriminatory bias of algorithms and AI used in various clinical applications (e.g. Panch et al., 2019; Grote & Keeling, 2022; Ricci Lara et al., 2022; Yang et al. 2024).

For instance, predictive models for in-hospital mortality and other clinically relevant parameters have been found to show insufficient performance for minority populations (Röösli et al., 2022). Moreover, AI models can predict sensitive → attributes such as race, gender based on scanning images, pointing to engrained discriminatory biases (Vyas et al., 2020; Korot et al., 2021; Howard et al., 2021). Vorisek and colleagues have conducted a survey about clinical AI practitioners on how they rate fairness of clinically deployed AI systems; the outcome was that a majority views these tools as only “moderately fair”. Lack of knowledge was identified by 45% of respondents as a source of bias (Vorisek et al., 2023). This indicates that **more efforts may be needed to raise this topic among the community** (→ AI practitioners) and **raise awareness of how to address it**.

A pro-active attitude of looking out for bias sources should become routine when developing and deploying AI for health.

### **Monitoring discriminatory bias – ex ante and ex post:**

Discriminatory bias should be monitored

- a) in a **proactive manner from the very beginning of the → life cycle of AI in health** and → **AI evidence pathway for health**, e.g. by carefully considering hidden bias in data, conceptual

and scientific or clinical assumptions, beliefs, traditions (→ conceptual relevance; → valid clinical association / scientific validity), by ensuring that data compilations are bias free ([Ganapathi et al., 2022](#)), by collaborating closely with data providers in regard to their tools and metrics for discriminatory bias detection and by critically interrogating decisions along the → algorithm-to-model-transition, e.g. → performance metrics or → model calibration in regard to whether they may introduce discriminatory bias.

- b) **throughout the life cycle and evidence pathway of an AI system**, i.e. when gathering information on performance at pre- and post-deployment stages (see → avoiding discrimination and discriminatory bias).

There are tools and approaches for identifying discriminatory → algorithmic bias in health and medicine. These approaches should be applied however with the fact in mind that **not all discriminatory bias is due to algorithmic bias**.

Thus, safeguarding only against for → algorithmic bias is not sufficient for a complex and high-risk application field such as health and medicine (→ Monitoring and mitigation of possible discrimination throughout the evidence pathway).

#### **Methods for detecting and addressing unfair bias:**

We point to some relevant literature in the context of detection, removal and metrics of unfair bias:

- Proposals for **addressing unfair bias for health equity** have been made by Abràmoff and colleagues ([Abràmoff et al., 2023b](#)).
- **Methods for addressing unfair bias** have been reviewed by [Yang and colleagues \(2024\)](#).
- Issues of **fairness and bias specific to clinical application fields** such as image analysis or critical care have been discussed ([Xu et al., 2022](#); [Charpignon, 2023](#)).
- On a more **general level of AI model building, fairness and bias in machine learning** have been examined (e.g. [Caton & Haas, 2020](#)), with proposals for **removing discrimination for classification outcomes** ([Kamiran & Calders, 2009, 2012](#); [Kamishima et al., 2012](#)), to ensure equality in supervised learning ([Hardt et al., 2016](#)) and also for **certifying and removing disparate impact or discriminatory bias** ([Feldman et al., 2014](#)).
- Finally, there are many websites that offer views and tools including “**fairness metrics**” (e.g. [Shelf, 2024](#)).

- **Who may be concerned?** AI developers (including data scientists, modellers, life scientists and biomedical scientists advising modellers), AI deployers (e.g. in hospitals and other healthcare settings), clinicians, health information professionals.

#### **Monitoring and mitigation of possible discrimination throughout the evidence pathway**

**Parent terms:** Fairness- Avoiding discrimination: inclusion, diversity and plurality

It is critical that AI systems are monitored for discriminatory bias **throughout the life cycle and evidence pathway** (→ life cycle of AI in health; → AI evidence pathway for health). Information on performance at pre- and post-deployment stages should be performed with the possibility in mind that **performance issues** may not be due only to insufficient training, or → **distributional drift /shift**, but maybe due to **discriminatory bias which can manifest itself also as a safety issue**.

Possible indicators are the same as for other factors (e.g. drift) and it is **not trivial to disentangle potential sources of performance degradation from genuine problems with discriminatory → algorithmic bias**. A clear indicator for discriminatory bias is lack of predictive power for specific groups, detectable through lower → **accuracy**, lower → **precision**, or lower reliability (e.g. → **replicability**, → **reproducibility**) when analysing datasets relating to such groups.

Collecting such data has however **potential privacy implications** (→ PRIVACY PROTECTION), requiring consent and proportionality between data collected and analytical goals and having wider social implications, including against the background of historical experiences made by minority groups concerning their health data collection ([Mittelstadt, 2021](#); see section “collection of sensitive data for bias and fairness auditing”). The EU’s AI Act outlines conditions specifically for data collection in regard to bias detection (Article 10.5; [EU, 2024a](#)).

From a procedural point of view, monitoring of discriminatory bias can be achieved through post-deployment monitoring, → clinical evaluation (pre and post-market stages), post-market surveillance activities (→ post-market surveillance, market surveillance, corrective action) and auditing (→ auditability and auditing) (see → RESPONSIBILITY). Successful monitoring requires however:

- **Sensitivity or ‘attunement’ to the possibility that performance issues are due to discriminatory bias** either engrained in the model (→ algorithm-to-model transition) or the manner in which the AI system is used in clinical practice.
- **Ensuring sufficient documentation** about relevant steps in the → algorithm-to-model transition, so as to allow → traceability and → auditability and auditing of possible sources of discriminatory bias that may have crept into a model. → Traceability is both an issue of → TRANSPARENCY but also of NON-MALEFICENCE. An audit will only be able to make connections between potential causal factors and discriminatory bias if there is properly recorded and traceable documentation available to auditors, concerning *inter alia* data, modelling decisions, scientific assumptions etc. (→ algorithm-to-model-transition).
- **Installing processes ensuring → Failure transparency**, i.e. the transparent and evidence-oriented gathering and documentation of situations, circumstances and conditions (e.g. → use context, → use environment), where an AI system showed performance issues that may point to discriminatory bias; this includes documentation of contextual information, e.g. on → clinical pathways and workflows of which the AI system was / is a component.

- **Who may be concerned?** AI developers (including data scientists, modellers, life scientists and biomedical scientists advising modellers), AI deployers (e.g. in hospitals and other healthcare settings), clinicians, health information professionals, AI auditors, compliance experts (e.g. working for third party bodies/notified bodies).

## Health equality & health equity

**Parent term:** Fairness

As discussed before, fairness in health must be seen also from the perspective of **health equality** and **health equity**. The Oviedo convention (→ DIGNITY, FREEDOM AND AUTONOMY) stipulates in Article 3 “equitable access to health care” as follows: “*Parties, taking into account health needs and available resources, shall take appropriate measures with a view to providing, within their jurisdiction, equitable access to health care of appropriate quality.*” We summarise both terms as follows:

- **Health equality** means that each individual or a specific group of people is treated in the same way. In the context of AI in health this means that AI outcomes do not discriminate (→ FAIRNESS, → bias) and resources as well as opportunities (e.g. for diagnostics, treatments) are the same for everybody (e.g. in a health system or a country, but also globally).
- **Health equity** is “*the absence of unfair, avoidable or remediable differences in health status among population groups defined socially, economically, demographically or geographically*” ([WHO, health promotion glossary of terms, 2021d; see also social determinants of health: WHO, 2024b](#)). In practice, health equity can mean that equal outcomes require with different means for different people, taking individual circumstances and needs into account ([Milken Institute School of Public Health, 2024](#)).

From a technical perspective, health equality and equity will require care in regard to the use of sensitive → attributes, e.g. when constructing a predictor under supervised (e.g. [Hardt et al., 2016](#)).

Given the prospect that the deployment of AI in healthcare will likely be very heterogeneous within and between countries, there is the danger that such “geographical bias” ([Mittelstadt, 2021](#)) will create equalities and inequities in regard to high quality medical care or, worse, that existing inequalities and inequities will be further exacerbated:

*“If AI systems raise the quality of care, for example by providing more accurate or efficient diagnosis, expanded access to care, or through the development of new pharmaceutical and therapeutic interventions, then patients served by ‘early adopter’ regions or health institutions will benefit before others.” ... “The inconsistent rollout of AI systems with uncertain impacts on access and care quality poses a risk of creating new health inequalities in member states. It may prove to be the case that regions that have historically faced unequal access or lower quality care are seen as key test beds for AI-mediated care. Patients in these areas may have better access to AI systems, such as chatbots or telemedicine, but continue to face limited access to human care or face-to-face clinical encounters. The likelihood of this risk depends largely on the strategic role given to AI systems. If they are treated as a potential replacement for face-to-face care, rather than as a means to free up clinicians’ time greater inequality in access to human care seems inevitable.”*

In some countries or global areas, the non-uniform (unequal) accessibility to healthcare constitutes an important barrier to both preventive and curative health services (e.g. disparity between rural and urban India: see [NITI Aayog, 2018](#)).

Note should be taken that the potential differences of AI uptake and way of implementation in national or regional health systems would not only have a bearing on individual and global health equity, but also on the patient-physician relationship and the manner in which healthcare may develop in the future given the possibilities for automation. See the paragraph on **Replacement of physical patient-physician interaction with virtual systems** under → DIGNITY, FREEDOM AND AUTONOMY – Upholding a trustful patient-physician relationship. Mentioned difference in roll-out of AI would also touch on the need for open democratic debates about the future of healthcare (see - >AI and the development of healthcare).

- **Who may be concerned?** AI developers, AI deployers (e.g. in hospitals and other healthcare settings), health policy makers, health system administrators, clinicians, researchers and policy specialists in the area of ethics, bioethics, global development.

## Unavoidable trade-offs

**Parent term:** Fairness

As discussed in more detail under → **intrinsic incompatibilities or ‘trade-offs’**, systems providing predictive classifications cannot satisfy different notions of fairness simultaneously (e.g. optimisation for positive group will bring an comparably unfair disadvantage to patients belonging to the negative group).

That means that there is inherently a trade-off between various requirements. This tension should be understood and carefully considered when designing and developing AI models for health applications. Decisions regarding trade-offs should be clearly described and documented, also in the interest of → failure transparency, → traceability and auditability (→ auditability and auditing), so as to support efficient analyses of potential discriminatory → algorithmic bias.

In addition, fairness in a more general sense requires also to balance competing interests of various stakeholders ([EU HLEG, 2019](#); [WHO ethics guidelines, 2021a](#)).

- **Who may be concerned?** AI developers (including data scientists, modellers, life scientists and biomedical scientists advising modellers), AI deployers (e.g. in hospitals and other healthcare settings),

clinicians, health information professionals, AI auditors, compliance experts (e.g. working for third party bodies/notified bodies).

## Universal versus targeted design

### Parent term: Fairness

While generally a so-called “universal design” of an AI system is desirable ([EU HLEG, 2019](#)), i.e. one that is applicable to the greatest number of individuals, there are situations where AI systems (e.g. diagnostic tools or algorithms for supporting clinical decision making) will need to be tailored to specific groups or patients.

Even for seemingly trivial tools like oxygen-measuring devices, a “targeted” design may be required due to differences in skin colour of various ethnicities.

Design choices of universal versus targeted design should be justified using scientific and clinical evidence and documented.

- **Who may be concerned?** AI developers (including data scientists, modellers, life scientists and biomedical scientists advising modellers), AI deployers (e.g. in hospitals and other healthcare settings), clinicians.

## A.8 Solidarity

### Concept description

Solidarity is conceptually close to equality, non-discrimination and fairness (see for instance EU high-level expert group guidance, page 11; [EU, 2021](#)). However, fairness/equality and solidarity are not fully congruent.

The term solidarity is derived from the ancient Roman legal concept of 'in solidum' which referred to joint contractual obligations where possible liabilities were shared. There is a continuation of that notion in the modern meaning of solidarity which can be described as a **relationship of unity, mutual indebtedness within a group or society at large**.

Solidarity is based on **mutual benevolence and fellowship between people and groups**, with an understanding that **nobody should be left behind**, in particular **vulnerable individuals or groups**. Its notion of pro-active effort to provide for everyone is close to equality, but not exactly the same: equality refers to people *being treated* equally.

We distinguish five translational concepts under solidarity. The first two are health-specific (Nr. 1) or have a particularly strong relevance in health (Nr. 2) due to the complexity of the health value chain and the need for collaboration to harvest benefits of AI in the health domain ([OECD, 2024](#)). The other three concepts are not specific to health.

1. Most fundamentally, **solidarity is about maintaining open, democratic and solidary societies**, characterised by participatory approaches and public and sufficiently informed debate enabling a culture of dialogue and involving different groupings of a society. For AI in health, this broad societal dimension is mainly relevant for discussions concerning health system design and management, including the future evaluation of healthcare (→ **DIGNITY, FREEDOM AND AUTONOMY** → AI and the development of healthcare). While product safety is a fundamental pillar in this context, critical reflection and debate about how AI solutions are implemented in health systems and workflows should not be neglected.
2. **Solidarity with particularly vulnerable groups**: rare indications, orphan indications and mental health well-being
3. **Solidarity in terms of workplace, skills, training as well as structures and tools that enable collaboration** across the evidence pathway of AI.
4. **Solidarity in terms of shared benefits of technology** and enabling a range of actors (including SMEs, various global areas) to develop AI technology.
5. **Solidarity in terms of avoiding colonial and exploitative structures** related to AI development and deployment, e.g. when considering the global north-south divide.

### Explanatory note

Solidarity is the least frequently used of the 11 most commonly used ethical principles according ([Jobin et al, 2019](#)). Its main notions are '*social cohesion*' and '*social security*'. As indicated, solidarity is often mentioned in the context of fairness and equality (e.g. [EU HLEG, 2019](#)). Solidarity relates to → **health equality and health equity** (→ **FAIRNESS**).

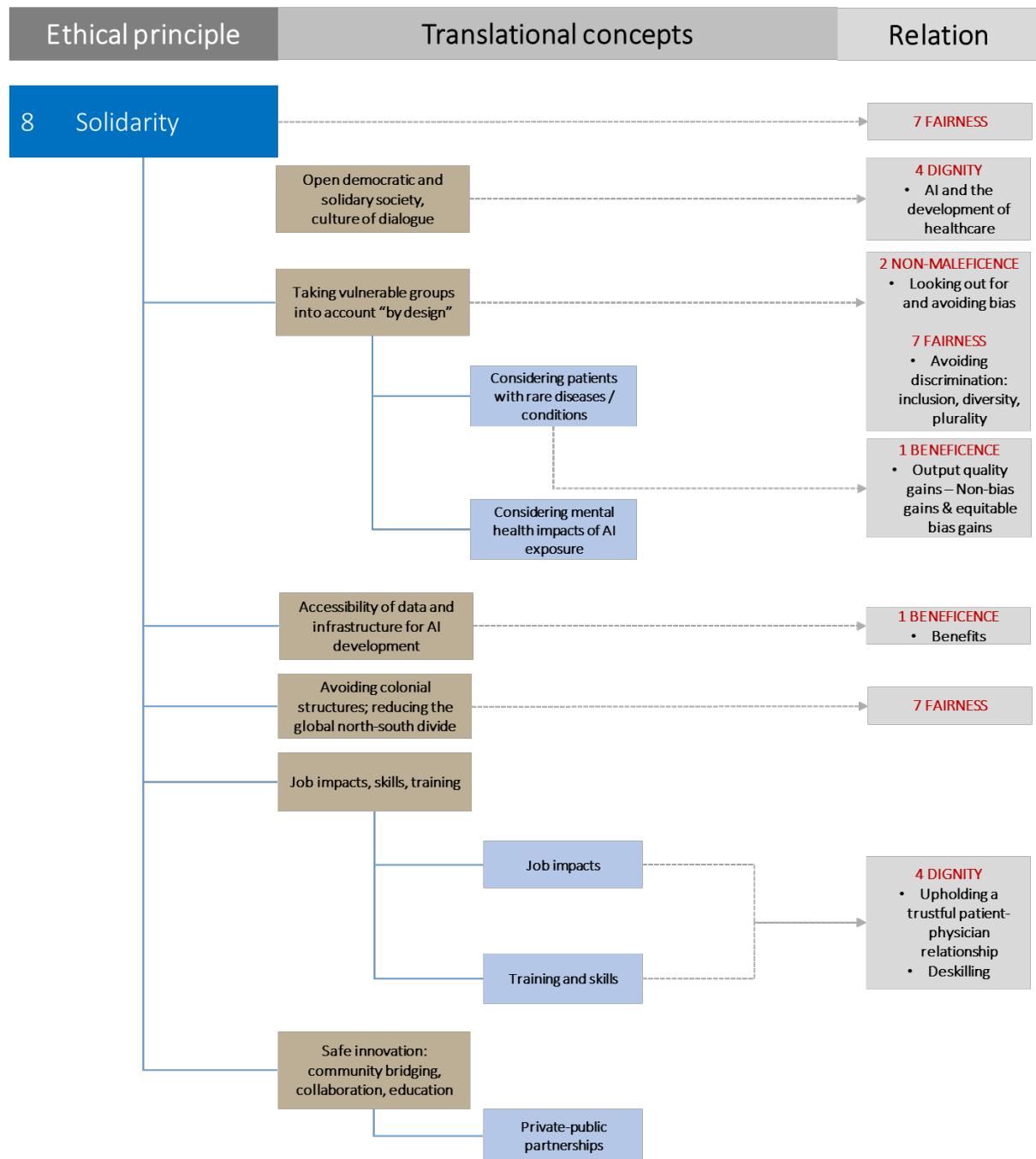
### Term relationship

Related terms:

- FAIRNESS
- Health equality and health equity
- DIGNITY, FREEDOM AND AUTONOMY
- Patient primacy
- Human primacy
- Intelligibility

- Interpretability and explainability

Ontological organisation of the ethical principle of “solidarity” and its translational concepts. Relations to other translational concepts are indicated by grey / stippled arrows.



## Solidarity: translational concepts

### Open, democratic and solidary society, culture of dialogue

Most fundamentally, **solidarity is about maintaining open, democratic and solidary societies**, characterised by **participatory approaches and public and sufficiently informed debate enabling a culture of dialogue** involving different groupings of a society.

For AI in health, this broad societal dimension is mainly relevant for discussions concerning **health system design, management and development, including the future evaluation of healthcare** (see → DIGNITY, FREEDOM AND AUTONOMY; → AI and the development of healthcare).

Technological developments linked to, inter alia, (big) health data, wearable devices, internet of things and an increasing connectedness of medical technologies, devices and systems as well as the rapid development in regard to large multi-modal models and their use in health care will require a constant readjustment of discussions of benefits versus risks of these technologies. Importantly, this will not only concern the make-up of these technologies and their 'intrinsic' risks or safety, but – perhaps more importantly – the way they are implemented and used at a larger scale.

In Europe, there is an obligation to all parties that are signatories of the Oviedo Convention ([Council of Europe, 1997](#)), to maintain adequate public debate about developments in biomedicine and associated social, ethical, medial, economic and legal implications. Article 28 states: "*Parties to this Convention shall see to it that the fundamental questions raised by the developments of biology and medicine are the subject of appropriate public discussion in the light, in particular, of relevant medical, social, economic, ethical and legal implications, and that their possible application is made the subject of appropriate consultation.*"

### Taking vulnerable groups into account "by design"

#### **Parent term:** Solidarity

While discrimination, exclusion and inequality (i.e. lack of fairness) refer to the unintentional or intentional **exclusion** of specific people and vulnerable groups from benefits, including those furnished by an AI system, solidarity is understood as a desideratum to avoid, even unintentionally, **leaving groups or individuals behind** in regard to the shared benefits of AI technologies in health.

AI system development should not exclusively focus on the majority of patients and medical indications but consider also **groups and patients with indications that may currently be disadvantaged from a perspective of not being provided with technologies that might improve their specific situation, condition or help diagnosing and treating their disease**. N.B. This specific disadvantage refers to being 'left behind' (a matter of solidarity) and not 'excluded' or 'discriminated' by design features of available (AI) technologies. This would be a matter of → FAIRNESS.

It is helpful in this context to consider the issue of **orphan drugs and orphan medical devices**, i.e. drugs or devices that are not developed by industry although there would be a public health need; similarly, orphan indications are those indications for which there is no medicinal product or device (for diagnosing or treating the disease) ([Orphanet, 2024](#)).

Often such indications are associated with **rare diseases**; orphan health products are not produced due to economic considerations based on the high cost of developing and bringing health products to the market that, by the very nature of their target indication, would have a small market share. Similarly, **AI systems may focus on indications that are economically viable**, leaving behind specific indications and hence people.

## Considering patients with rare diseases / conditions

**Parent term:** Solidarity – Taking vulnerable groups into account “by design”

AI has great potential to support the diagnosis and treatment of patients suffering from so-called “rare diseases”, characterised by their low prevalence and diagnostic complexities ([Abdallah et al., 2023](#)).

Early diagnosis of genetic rare diseases is complicated by genetic and clinical heterogeneity of patients, resulting in major challenges concerning the availability of diagnostic tools that cover the wide range of factors (e.g. biomarkers) underlying specific forms of rare diseases.

AI systems with their ability to analyse large data sets efficiently and quickly may help tackling the necessary suite of tools required to address rare diseases, including in vitro diagnostic medical devices, image analysis, drug development and repurposing.

Consideration should be given **to address the needs of patients with rare diseases** when developing various AI tools, ranging from health research to → AI-enabled medical device software.

## Considering mental health impacts of AI exposure

**Parent term:** Solidarity – Taking vulnerable groups into account “by design”

Another element of solidarity relates to the impacts of AI technologies on the **mental health and well-being** of people and in particular of vulnerable individuals. There is an ongoing debate on the potential detrimental effects of AI technologies on mental health (e.g. [Huang et al., 2024b](#); [Ettman & Galea, 2023](#)).

Clearly, technologies that are designed to interact with natural persons (e.g. → **conversational agents**, chatbots) should be examined in terms of their possible impact on the mental well-being of users. This is obviously not only a question of technology (e.g. the technological realisation of → **affective computing**), but probably more so a matter of *how such systems are implemented* and whether they might lead to a feeling of *alienation, helplessness and mechanisation* by substituting direct interactions between people.

## Accessibility of data and infrastructure for AI development

**Parent term:** Solidarity

An important aspect of solidarity in the context of AI in health concerns the **accessibility to high-quality data for AI development** that are inclusive and non-discriminatory ([Griesinger et al., 2022](#)).

The future of health (FOH), a community of senior health leaders has identified various obstacles to the successful implementation of AI in health, including accessibility to health data, proposing to design mechanisms for sharing relevant data ([Silcox et al., 2024](#)). Please see in this context also the EU Commission’s website on data governance with a focus on mechanisms and standards for sharing data ([European Commission, 2020b](#)).

Equally, there may be infrastructure barriers regarding the ability of various actors to develop AI. This includes aspects of sufficient computing power, data storage capacity, data exchange mechanisms etc.

Accessibility to data and infrastructure is important to avoid monopoly situations or dominance of the AI-enabled health product market by big companies. Enabling smaller actors and organisations in the global south to develop AI will also help tailoring AI-enabled systems to local needs and help integrating these into existing workflows, → **use environments** and → **use contexts**.

## Avoiding colonial structures; reducing the global north-south divide

**Parent term:** Solidarity

### **Data colonialism**

So-called “data colonialism” (see for instance: [Gray, 2023](#); [Purdue University, 2024](#)) refers to the situation that big organisations and big corporations may claim ownership of and privatize data that have been produced and are belonging to individuals whose rights concerning consent of data use, privacy and auton-

omy (e.g. withdrawal of consent to use data) are not or not sufficiently respected. Data colonialism reflects also stark global differences in statistical capacity and hence possibility to use data for AI training. In general, there is a pronounced north-south divide with a skew of valorisation towards the global north ([World Economic Forum, 2023](#)).

There is therefore a risk of a growing divide between global regions and actors who collect, accumulate, buy, analyse and govern large set of biomedical data often collected from underrepresented groups and those that provide the data but remain with little control over it – in particular in the case of low- and middle-income countries ([WHO, 2021a; Muldoon & Wu, 2023; Arora et al., 2023; Okolo, 2024](#)).

### **Digital colonisation**

Another facet of this issue is “digital colonisation”, i.e. a new colonisation of the global south if foreign companies (largely from the global north) continue to feed on African data for AI development without involving local actors ([United Nations, 2024d, f; Roche et al., 2021](#)).

Digital colonial structures concern also ‘supply chain’ aspects of data. AI and tech companies have outsourced data labelling or data annotation (in order to render the data usable for supervised → machine learning) to the global south where people label data sets for various applications (e.g. driverless cars) often under precarious conditions, without liveable wage or proper work contracts (see for instance [Sambasivan & Veeraraghavan 2022; MIT technology review, 2024; DataEthics EU, 2024; Access partnership, 2024](#)).

## Job impacts, skills, training

### **Parent term:** Solidarity

#### Job impacts

### **Parent terms:** Solidarity– Jobs, skills and community bridging

There is ample literature on anticipated disruptions in the workplace due to the increasing use of AI systems and their potential to support or even replace human agency for specific tasks. Note that WHO considers this under the umbrella term of ‘workplace sustainability’ ([WHO, 2021a](#)).

In the health sector, this has positive sides given the global shortage of qualified healthcare professionals with AI having the potential to mitigate the effects of this shortage (see for instance [NITI Aayog of India, 2018](#)), but also negative sides concerning potential job losses of qualified staff.

In their seminal study on the impact of computerisation including AI on jobs, Frey and Osborne used a quantitative approach to predict how likely a wide range of professions and jobs will be affected by the use of digital technologies ([Frey & Osborne, 2017](#)). Regarding healthcare, the study found that while more routine jobs like medical records management, medical transcriptionists, health information technicians and medical secretaries carried a high potential for computerisation, also physicians had a likelihood of 42% of their profession being automated. A recent survey found that healthcare workers are indeed concerned about job security but also the impact of AI on patient care ([Rony et al., 2024](#)).

## Training and skills

### **Parent terms:** Solidarity– Jobs, skills and community bridging

As pointed out by the WHO ([WHO, 2021](#)), responsible AI in healthcare requires that potential disruptions in the workplace are factored in when rolling out AI in healthcare and medicine (the report refers to “workplace sustainability”). With an increasing consensus that AI in health should be used to support human actors, especially when it comes to the patient-physician relationship, a focus needs to be on **training programmes for healthcare workers** to adapt to the responsible use of AI systems in their workplace. Ideally, the efficiency or productivity gains of AI in healthcare settings ([Abràmoff et al., 2023a](#)) (see also → **BENEFICENCE**) should be translated in improved focus and more time for patient care by healthcare professionals.

Similarly, it would seem necessary that automated systems structuring, managing and translating medical information in electronic health records require maintenance and supervision by sufficiently trained **health information professionals**. The same holds for using AI systems including large language models for exploiting big health data for basic biomedical research and clinical research as well as, possibly, for clinical decision-making: such systems will require technical support and properly trained users (e.g. researchers, physicians, biomedical experts) that understand the functioning and limitations of such models and are able to prompt systems in the best possible way and can use results with prudence and circumspection.

Further, the increasing digitisation of health including through the uptake of AI systems enhances the vulnerability of health settings to cybersecurity attacks with incidents affecting also patients' health ([Reina & Griesinger, 2024a](#)). To counter these risks and enhance the cyber resilience of health systems and healthcare settings, properly trained IT system administrators and cybersecurity specialists working in health settings are required (see also EU political guidelines of the new Commission, 2024, page 9; [European Commission, 2024](#)). Thus, the accelerating uptake of AI may also create additional job opportunities in health.

## Safe innovation: community bridging, collaboration, education

### **Parent term:** Solidarity

#### **Safe innovation**

Innovation is about introducing new concepts, approaches, processes and procedures. Inherently novelty is associated with risks – newness means a lack of experience with potential pitfalls and hazards.

The concept of 'safe innovation' is about finding the sweet spot between sufficient *ex ante* characterisation and mitigation of avoidable risks to ensure maximum attainable safety (→ **AI safety**) without however negatively affecting potential benefits to a degree that would degrade the innovation's → **added value** to so great an extent as to render the innovation pointless. Thus, safe innovation is about determining what level of *unavoidable risks* (see → **NON-MALEFICENCE**) are acceptable given the *potential* benefits (→ **BENEFICENCE**) of the innovation. Relevant legislations in the health area (e.g. the EU medical devices Regulations; [EU, 2017a, b](#)) are built on the concept of the so-called "benefit-risk ratio" approach<sup>42</sup>. Clearly, to determine the acceptability of risks, robust evidence on the claimed benefits and possible risks is needed (→ see **BENEFICENCE**). Safe innovation thus encapsulates the determination of the *kind and level of evidence* required to make informed decisions in a process that is essentially a negotiation between different parties with overlapping but not identical interests.

#### **Collaboration for evidence: the evidence pathway of AI in health**

Given the complexity of the life cycle and value chain of AI in health, open debate and a collaborative attitude of the diverse communities (→ **AI actors and communities**) is critical – it is in fact a matter of **solidarity** vis-à-vis stakeholders and patients. In its report on responsible AI in health, the **OECD** ([OECD, 2024d](#)) has outlined key areas for collective action: trust, capacity building, evaluation, and collaboration. This recognises that the primary forces that are needed to unlock the value from artificial intelligence are people-based and not technical.

We argue that the safe innovation aspect of AI governance needs to pay more attention to negotiating required types and levels of evidence. This includes notably evidence from → **model evaluation** and → **clinical validation**, that create information, albeit limited, from situations that are closer to real-world use conditions as opposed to research settings and highly optimised conditions and that take relevant → **use**

<sup>42</sup> The AI Act ([EU, 2024a](#)) does not use a benefit-risk approach, but instead matches the level of regulation required to risk strata (for a comment see [Fraser et al., 2024](#)). The EU AI Act points to more horizontal sectoral legislations like the EU medical devices and in vitro diagnostic medical devices Regulations ([EU, 2017a,b](#)) which use the concept of benefit-risk ratio and risk acceptability.

contexts and → use environments into account. Both are indispensable for an accurate estimation of risks but also realistic benefits.

### ***AI evidence pathway and this ontology as tools to support community-bridging, literacy and education***

We propose the “→ AI evidence pathway for health” (see also section 1.2) as a framework for bridging the relevant communities in view of collaborative identification of evidence needs and evaluation processes to ultimately support safe innovation, development, deployment and use of AI solutions in health.

This ontology is a key element of the pathway: by providing an explication of relevant ethical and translational as well as fundamental concepts (from machine learning to clinical use), it will enable communities to collaborate more effectively with each other, understand key concepts that may be alien to actors from other fields and to help bridging specialist language.

Sufficient ‘**literacy**’ of relevant fields (ethical, technical, clinical) is an important prerequisite for effective collaboration of multidisciplinary actors during development, deployment and monitoring as well as for the safe and secure use of AI in real-world settings. The current ontology is intended to support such literacy. Sufficient literacy should encompass all actors of the AI evidence pathway:

- ***Literacy in health and healthcare:***

- While data scientists, AI designers and machine-learning specialists engaged in building AI solutions for health should have a minimum of ‘**literacy in health and healthcare**’, the current focus is understandably on AI literacy.

- ***AI literacy:***

- Medical training: The parliamentary assembly of the [Council of Europe \(2020a\)](#) called on Member States to “adapt their education and training systems to integrate **AI literacy** into the curricula of schools and medical training institutions, with an emphasis on the ethical principles of AI and responsible uses of AI applications.”

This would ensure that downstream actors such as clinicians, healthcare professionals, medical technicians and health information professionals (HIPs) (see → AI value chain actors) can use AI in a responsible and informed manner, knowing about its basic functioning, applicability and, most importantly, its limitations. To use AI effectively, safely and ethically, health professionals need a range of skills such as a) “AI fundamentals” (e.g. a general understanding of models, neural networks), b) ethical and legal aspects, followed by a range of topics, notably evaluation, validation, biases etc. ([Gazquez-Garcia et al., 2025](#)). Also, regulators and health technology assessors require sufficient “AI literacy” for effective governance and evaluation of AI tools.

- AI system providers and deployers: Article 4 of the EU’s AI Act ([EU 2024a](#)) stipulates that AI system providers and deployers must ensure a sufficient level of **AI literacy** of their staff and “other persons dealing with the operation and use of AI systems on their behalf...”. A FAQ document on AI literacy as required by Article 4 has been published ([European Commission, 2025d](#)).

### Private-public partnerships

**Parent terms:** Solidarity– Safe innovation: community bridging, collaboration, education

In the context of AI, private-public partnerships could support safe innovation as a way to tackle technological, social and regulatory challenges. Such collaborations may be in the interest of both partners. Surprisingly, the major public guidelines on AI (ethical) principles do not explicitly refer to private-public partnerships ([Hagendorff, 2021](#)).

Private-public partnership could be used for various purposes where there is a common interest of the public and private sphere:

### ***Exchange of views, knowledge, insights, development of suitable frameworks and guidance***

Partnerships could serve to

- exchange of knowledge, views and insights, e.g. on technological aspects or regulatory or legal requirements
- ensure understanding of (novel) technologies and developments in the dynamic field of digitisation, AI and data
- conduct joint assessment of new technological developments and associated potential benefits, challenges, and risks, e.g. novel AI-types (→ AI typology) or → AI techniques with particular risks (e.g. → foundation models; generative AI; → data modality) or → synthetic health data.
- define evidence needs especially for sensitive and high-risk applications.
- accelerate the development of assessment frameworks, methods and metrics or guidance that are fit for purpose regarding regulatory requirements while keeping an eye on unnecessary burdens for industry in view of competitiveness and innovation capacity

### ***Reskilling***

Private public partnerships are also discussed in the context of ensuring that people acquire the necessary skills required for a successful and safe rollout of AI solutions ('reskilling') ([World Economic Forum, 2024](#)). Private public partnerships could help defining, elaborating skills and skill levels required and elaborate recommendations for training on technology-related topics, ethical evaluation, impact assessments and regulatory matters.

### ***Regulatory sandboxes***

'Regulatory sandboxes' can be considered a specific form of private- public partnership, particularly important for smaller companies or start-ups. Sandboxes may help, in a confined and supervised environment involving regulatory authorities, to advance the safe development, training and validation of AI systems prior to their placing on the market. The EU's AI Act (Article 57; [EU, 2024a](#)) provides for 'sandboxes' ([EU, 2024](#)).

### ***Academic research***

Academic research is another area of private-public partnership, where corporate sponsor support for academic research that is also in their interest. While this may promote research activities, there are also concerns in regard to safeguarding the freedom of research ([Hagendorff, 2020](#)) in particular in situations where public funding of independent research is de facto replaced by funds from private actors, that may (or may be perceived to) influence the direction and scope of the research including interpretation and presentation of research results.

## A.9 Sustainability

### Concept description

AI development and use should take into consideration the sustainability principle, i.e. the protection of the environmental and the planetary resources in a manner compatible with sustaining future generations. The overall economic management and development associated with AI as a tool of potential significant benefit needs to be done in a manner that meets “*...the needs of the present without compromising the ability of future generations to meet their own needs.*” ([UN, 1987: Brundtland commission report](#); [UN sustainability website](#), [UN, 2024a](#)).

AI development and/or operation may require large amounts of energy for powering computers running → **machine learning algorithms** during training and use and for powering data servers. AI has a large environmental footprint, in particular if the required electrical energy is produced by fossil energy sources. The WHO has highlighted that AI technologies used for health purposes need to be aligned with the goal of health systems being sustainable, i.e. the way in which healthcare is organised, planned, managed and delivered ([WHO, 2021a](#)). Sustainability has been high on agenda of the 2025 Paris AI Action Summit to which more than 100 countries participated ([French Government, 2025](#)).

The growing employment of AI for various purposes including health and healthcare needs thus to be considered in the context of urgent global challenges such as the increased amount of atmospheric CO<sub>2</sub> and resulting global warming. The UN sustainable development goals are an agreed basic framework for improving the lives of populations around the world and mitigating the effects of climate change ([UN, 2024b](#)).

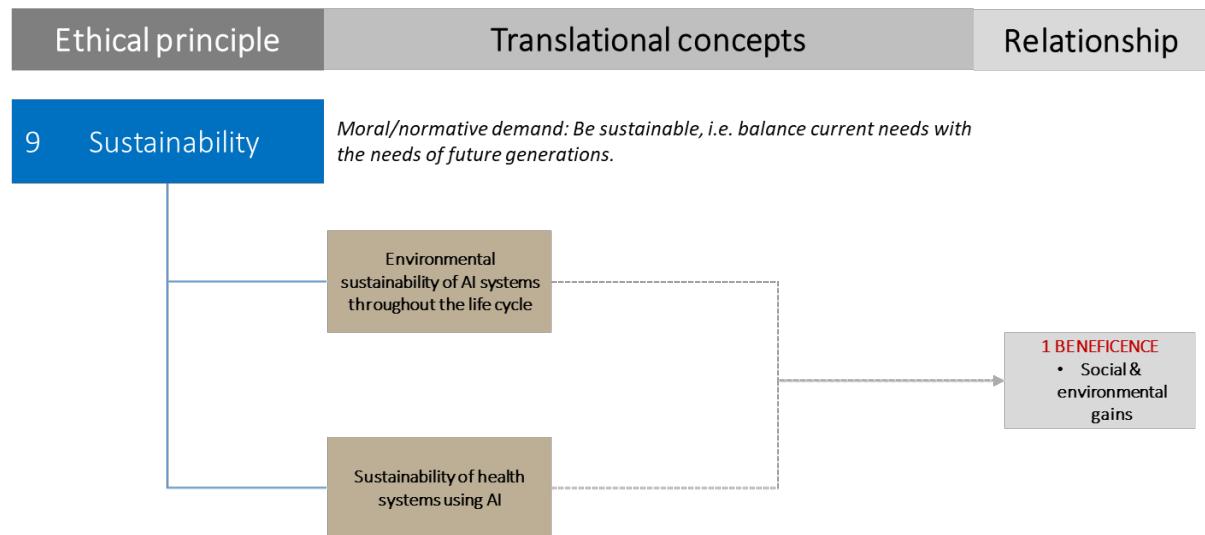
### Explanatory note

Occasionally sustainability is used to denote not only environmental sustainability, but also sustainability regarding socioeconomic aspects, such as the “sustainability of work places” ([WHO, 2021a](#)). We use → **SOLIDARITY** for effects of AI on the workplace. We hence restrict sustainability to the original environmental notion of the word.

### Term relationship

N.A.

Ontological organisation of the ethical principle of “sustainability” and its translational concepts. Relations to other translational concepts are indicated.



## Sustainability: translational concepts

### Environmental sustainability of AI system throughout the life cycle

#### **Parent term:** Sustainability

While politicians have an important responsibility to create conditions of energy production that are sustainable so that AI will be sustainable to all populations globally, actors developing AI have a more immediate responsibility to reduce the energy consumption of AI throughout the lifecycle and especially during training. In particular, deep neural networks possess a number of network parameters (e.g. weights to inputs) which can enhance accuracy but increases the number of operations during inference.

Energy consumption can be reduced through various measures ([Baker, 2024](#)). These include model pruning, use of processors that consume less electricity, integration of processing into storage, thus reducing data centre needs, distributed computing and AI workload management and task scheduling.

Energy consumption of → federated learning & split learning can be reduced by scheduling communication resources of IoT networks ([De Oliveira et al., 2024](#)).

So-called “early stopping”, a measure to identify lowest possible training epoch number without overfitting, is a way of lowering energy consumption ([Requero et al., 2025](#)); this involves monitoring → model performance during training (e.g. using a validation set) and stopping it as soon as the performance starts to degrade.

Enhancing the intrinsic energy efficiency of neural networks is an intense area of research. Specific network design choices ([Tripp et al., 2024](#)) or algorithmic control of training can result in an acceptable trade-off between performance and lower energy consumption of deep neural networks ([Lazzaro et al., 2022](#)).

Modular use of task-complementary small networks may enhance energy efficiency compared to using one large energy-hungry network ([Kinnas et al., 2024](#)).

Novel AI techniques utilizing for instance spiking neural networks may help reducing energy consumption ([Li et al., 2022; Islam et al., 2024](#)).

The machine-learning and data science platform “hugging face” has recently published an “AI energy score” to “*to establish comparable energy efficiency ratings for AI models, helping the industry make informed decisions about sustainability in AI development.*” ([HuggingFace, 2025](#))

### Sustainability of health systems using AI

#### **Parent term:** Sustainability

When adopting and integrating AI systems in health systems, relevant decision makers need to consider the sustainability of AI systems. The potential benefits of the AI system (→ **BENEFICENCE**) must be weighed against and negative impacts on sustainability of its use.

## 4 Ontology B: Fundamental concepts

### B.1 AI and AI systems

#### Artificial intelligence (AI)

**Cluster:** B.1 AI and AI systems

##### Concept description

The term artificial intelligence (AI) refers to two areas: → AI as a scientific field and → AI systems, i.e. AI-based technological tools that fulfil tasks that would require natural intelligence if done by humans or animals. Such systems are commonly referred to as "AI".

The two areas are strongly linked: some → AI techniques relate to neuroscientific findings and, in a broader context, the → computational theory of mind (e.g. → artificial neural networks, → deep-learning, brain circuitry-inspired learning, e.g. [Shi et al., 2025](#)).

##### **A note of caution regarding the misleading use of the term 'intelligence'**

Although some AI techniques (→ artificial neural networks) are inspired by neuronal connections in neuronal (e.g. cortical) ensembles, it is a common misconception that AI produces its 'intelligent' outputs in the *same way as a neuronal cell ensemble, or a brain*: while AI may 'mimic' human or natural intelligence as judged by its *outputs*, it does not necessarily *recapitulate* the biological mechanisms underlying natural intelligence: the fact that 'outputs' match, does not imply necessarily that the underlying processes are the same.

This has important repercussions for understanding and ultimately controlling AI: "*AI produced behaviour is alien, that is, it can fail in unexpected ways*" ([Weld & Bansal, 2019](#)). As an example, the fact that an AI system produces the same output as a human (e.g. a clinician or other human rater assessing radiological images), cannot per se be taken as an indication that an AI system looks out for the *same* features as an intelligent and sufficiently trained person. The AI may have learned irrelevant visual → **features** (so-called → **shortcut learning**; [Geirhos et al., 2020](#)). Thus, interpretability and, where this is not an intrinsic property of an AI model, its explainability are key for ensuring that an AI system produces a right output for the right reasons (→ **interpretability and explainability**).

##### Explanatory note

The classic definition of McCarthy, Minsky, Rochester and Shannon ([McCarthy et al., 1955](#)) in their proposal for the Dartmouth summer research project on artificial intelligence still describes AI tools of today – irrespective of the → AI technique:

*"...the artificial intelligence problem is taken to be that of making a machine behave in ways that would be called intelligent if a human were so behaving."*

Since coinage of the term in the mid-20<sup>th</sup> century, countless definitions for AI have been proposed. Many do not distinguish between the science of AI and AI-based technological tools/AI systems: AI can indeed be understood as both the science and practice of building machines that fulfil tasks that normally would require human (or biological/animal) intelligence/cognition.

Many definitions focus exclusively on AI systems or AI-based technology without however being explicit about the double notion of the term, i.e. 1) AI as a scientific "programme" (to elucidate the workings of the human mind through simulation) and 2) AI technology, as a hypotheses-testing simulation tool for working on the first. Discussing the technological application of a scientific field without exploring at least the basics of the scientific field risks a confused discourse both in scientific / expert circles as well as among laymen.

Further, many of these definitions are highly technical which may constrain their openness to novel features of AI solutions (see → **types of AI**).

## Term relationship:

Related terms:

- AI as a scientific field
- AI systems
- Computational theory of mind (CTM)
- Connectionism
- Functionalism

## AI as a scientific field

### Cluster: B.1 AI and AI systems

#### Concept description

The scientific field of artificial intelligence has two main interrelated goals:

- 1) To construct artificial systems (machines, "AI systems") by employing computer programs (→ **algorithms**) which, simply put, fulfil tasks which would otherwise require natural intelligence (e.g. of a human). This relates to other (more established) fields, notably data science, statistical modelling, neuroscience (and its insights in learning and memory, the functioning of cell ensembles with self-organising feature-detection properties (e.g. visual cortex functional maps; cortico-hippocampal circuits), cognitive science, philosophy, linguistics and others.
- 2) To explore an assumed fundamental nature of intelligence, thought to be independent from a specific **physical substrate** (e.g. neuronal networks in brains), including through the construction of artificial systems that mimic natural systems (modular organisation of neurons, neuronal connections and their strengthening according to cellular 'learning rules' or cell-level substrates of learning and memory) and by analysing the behaviour of these artefacts.

*Concept description based mainly on: Minsky, 1969; Koch & Poggio, 1987; Feldman, 2001; Scarcello (2019).*

#### Explanatory note

The first goal of AI can be seen as a branch of computer science and is based on, *inter alia*, mathematics, philosophy and logic, statistics, control theory, cybernetics and neurosciences.

The second goal is closely related to cognitive science, on which AI draws and in which it has participation. Cognitive science is an interdisciplinary science aiming at modelling and understanding human intelligence, drawing on the findings and methods of a range of disciplines, including philosophy, neurosciences, linguistics, psychology, sociology, artificial intelligence research.

Thus, while the methodologies of cognitive science and AI overlap to a significant extent, it is important to acknowledge that their goals are different. Cognitive science is mainly concerned with *understanding mental and cognitive phenomena* and reducing these to its constituting fundamental principles; AI in contrast has, in particular during the last few decades, an *application focus*: constructing systems that behave as if they had cognition, with the ultimate programmatic aim of building systems that can 'act' intelligently. The original aspiration of 'strong' AI is to 'reproduce' in an artificial system cognitive functions, mental states and consciousness, i.e. qualities that are bound to natural / human brains. This goal has shifted over time since AI does not necessarily "reproduce" human problem-solving capacities. Instead AI may tackle problems in another manner as neuronal ensembles in biological brains would.

Nevertheless, the most striking example for goal two is the now widespread use of artificial / virtual neural networks that are based on and/or inspired by findings and theories in the neurosciences and cognitive sciences, notably that of cell ensembles, feature detectors or neuronal substrates for learning and memory, e.g. changing inter-neuronal connectivity strength, Hebbian learning, synaptic long-term potentiation (LTP), synaptic/dendritic remodelling.

It is noteworthy that the design principles of the most powerful AI algorithms are inspired by neuronal networks in brains (i.e. AI neural networks exhibiting 'deep learning'), and specifically the *brain cortex*. Since many of neuroscience's insights about cortical function are derived from animal models (e.g. rodents, cats, macaque monkeys) and since the architecture of → **artificial neural networks** with its layers does neither truly reproduce brain connectivity nor plasticity, we suggest caution in regard to bold statements that AI systems 'recapitulate' human brains or human cortical function. Currently large deep-learning models are rather an incomplete *in-silico* modelling or recapitulation of smaller, local neuronal ensembles in the brain cortex.

#### Term relationship:

Related terms:

- AI as a scientific field
- AI typology
- AI technique
- AI-enabled medical device software

## Computational theory of mind (CTM)

### Cluster: B.1 AI and AI systems

#### Concept description

The theory that the computer can serve essentially as a model of the mind, which is understood as a 'computational system'. Computational 'functionalism' or simply 'functionalism'\* was introduced by the philosopher Hilary Putnam in the 1960ies and is a central philosophical concept of CTM.

Computational functionalism holds that *mental states* (thoughts, beliefs, desires etc.) that are accessible through introspection are identical with *computational states* of the brain and can hence be defined by computational processes and parameters as well as their association to biological 'inputs' and 'outputs'.

Consequently, functionalism *implies* that mimicking or recreating these computational processes can be done in other substrates than biological brains: the material 'substrate' is not decisive for the emergence of mind and mental states but the computational processes themselves realised in any appropriate material substrate (e.g. brain, semiconductor chips).

In its most consequential form, CTM posits that, in theory, artificial systems can be constructed that reproduce the brain's 'computational processes' that underlie or are identical with what we call 'mind'. Consequently, such systems would not only mimic mind or mental states but, in an ontological sense, have mind and its properties, e.g. cognition, consciousness, memory, intentionality.

The CTM debate fertilised and inspired cognitive sciences and the development of alternative approaches to the standard "Turing" computational model. For instance, probabilistic models of cognition modelling mental processes through algorithms, with neural processes being mere functional implementations of these algorithms (e.g. Griffiths et al., 2010). This 'top down' approach differs from the bottom-up 'connectionist' approach based primarily on neural mechanisms and the emergence of higher-level properties from neural network architecture and McCulloch Pitts neurons (→ **artificial neural networks**). Artificial neural networks have emerged as the most flexible and powerful way of implementing artificial intelligence at present (→ **foundation models (FM)**; → **generative AI**). Both approaches however are indebted to the concept of computational functionalism.

Importantly, the functionalist view was criticized by many philosophers, most vocally perhaps by John Searle who used the now widely popularised 'Chinese room' analogy as a critique of the remaining gap between mere functional states of a system and mental states that are characterised by representation, meaning and intentionality (Searle, 1984)\*\*. Notably, Putnam later became one of the most vocal critics of the 'identity' aspect of functionalism, without however abandoning the belief that mind and

consciousness are not necessarily bound to biological substrates such as neurons and brains ([Putnam, 1988](#)).

*\*) Hilary Putnam emphasizes that there are other conceivable descriptions for the concept captured by functionalism, e.g. a calculus-based notion of the mind or a computational notion of the mind ([Putnam, 1988](#)).*

*\*\*) For an introduction to Searle's argument, see Searle's Reith lectures ([Searle, 1984](#)) and Searle ([1980](#)).*

## Further reading

For an immensely informative short history of cognitive science and its relationship to various concepts of artificial intelligence, see [Varela F \(1992\)](#). Since the conceptual foundations of artificial intelligence were established decades ago, this short book is still highly relevant.

## Explanatory note

CTM is a grouping term for the foundational propositions underlying the technical claim and programme of a "strong" artificial intelligence, i.e. that it is possible to build machines that can think or, in its weaker formulation (see A. Turing), to construct machines whose "behaviour" would be indistinguishable from humans and which would be able to tackle a near infinite set of problems without additional training, currently discussed under the keyword "artificial general intelligence" (AGI) (e.g. [Goertzel, 2014](#)). Arbib ([1987](#)) has pointed out that the computer metaphor of the mind is a modern variation of the age-old question "is man a machine"? - with the famous example of Descartes' proposal that animals are automata but human beings have in addition a soul that communicates with the body-automaton via the pineal gland. Another example is the book ["L'homme machine"](#) by the materialist physician and philosopher de La Mettrie ([1747](#)).

In any case, CTM rests on key contributions in philosophy, science, mathematics and technology. Some key contributions include:

- Aristotle's conception of the soul
- Hobbe's concept of the mind as a "calculating machine"
- Lapicque's 'integrate and fire' model of describing the excitability of and input integration within neurons in the brain through biophysical membrane de- and repolarisation events ([Lapicque, 1907](#)).
- McCulloch and Pitts ([1943; reprinted 1990](#)) extended the physiological concept of Lapicque and reformulated it in a broader context by proposing that neurons in the brain can be *described by propositional logic* and each act like *miniature logical calculus operators*. Given their interconnectedness in an ensembles or brain 'circuits' (i.e. neuronal networks) this may result in highly complex brain operations.
- Norbert Wiener's cybernetic metaphor ("*Cybernetics, or Control and Communion in the Animal and Machine*", [Wiener, 1948](#)).
- Alan Turing's paper of 1950 aimed at replacing the question "can machines think" with a question of whether a human could be misled of thinking that a machine has consciousness based on the quality of its output\*
- Hilary Putnam's concept of functionalism (1960: in "Minds and machines" or "The nature of mental states"), meaning that mental states do not depend on the constitution or make-up of what produces that mental state but rather on the functioning of that make-up. Subsequently, Putnam advanced strong arguments against his own concept of (computational) functionalism.

Importantly, CTM gave rise in the later 20<sup>th</sup> century to *connectionism*, a fundamentally different model than the traditional Turing-style models. Connectionism is also inspired by neurobiology, albeit with a focus on → **artificial neural networks** composed of input layers, hidden layers and output layers of virtual neurons that are connected through weights either in a feed-forward or a recurrent (feed-forward and feed-backward way), thus recapitulating theories about functional cell ensembles (e.g. a column of primary visual cortex) in the brain. Connectionism has revolutionised machine learning. It draws not only on the concepts formalised in McCulloch and Pitts seminal paper but also on findings and theories of

neuroscience, such as Hebbian learning, variable synaptic strength and complex neuronal circuitry with extensive feedback loops (e.g. for memory formation).

*\*) Thus, strictly speaking, the Turing test does not aim at proving machine consciousness but is about examining the conditions under which a machine output becomes indistinguishable from human intelligence.* [Pinar Saygin et al \(2000\)](#) provide an overview over the various arguments concerning the Turing test.

#### Term relationship:

Related terms (not further elaborated in this ontology):

- Strong AI
- Functionalism
- Functional State Identity Theory (FSIT)
- Connectionism

## AI systems

### Cluster: B.1 AI and AI systems

#### Concept description

Physical systems based on the premises, research and technological application of the scientific field of → artificial intelligence and composed of a hardware substrate for operating an algorithm (often a → machine learning model, obtained through running data through a → machine learning algorithm) and, depending on the system, other hardware elements such as, for example, a user interface, sensors or actuators.

With 'AI-type' algorithm we refer to the broadest possible meaning of AI, i.e. algorithms that produce outputs that would typically require the intelligent agency of a human. This behaviouristic understanding of AI does not necessarily mean that the AI system is based on machine-learning or its subtype, so-called → deep learning in → artificial neural networks: for instance, AI includes also highly refined expert systems that are not based on machine learning. For more details see → AI technique.

In healthcare AI systems include combinations of medical device software (including → AI-enabled medical devices software) and relevant hardware.

#### Explanatory note

##### **Notions of the term ,AI system':**

- **For a general discussion of AI systems**, it is useful to connect the concept of 'AI system' to the research field of AI in order to emphasise the inextricable link between the two and to avoid a description that resorts to technical details which inevitably narrow down the scope of the mental concept, rendering it less in tune with progress in AI science.
- **In practical terms, AI systems are 'machines' or 'devices'** that run an algorithm for analysing data and producing an output that would, if produced by a human, require a specific capability of biological or human intelligence. 'Mimicking' *all* aspects of human intelligence, i.e. "artificial general intelligence" (AGI) remains so far elusive and will likely remain an illusion as long as AI systems are merely probabilistic input-output machines – as opposed to natural brains that represent the environment at any point in time in which "*an image of the macrocosm is set up and continuously adapted by learning*" ([Braitenberg & Schüz, 1991](#)).
- So-called 'intelligent' outputs of AI include classification of data, prediction of specific properties based on analysing representative real-world data, inferring properties, data-driven

- content ‘generation’ or synthetisation (→ generative AI), interaction with the environment via sensors and actuators.
- While AI systems may be inspired by neurobiological systems (e.g. neuronal networks in the brain cortex as a blue print for → artificial neural networks), it is a question of ongoing debate whether AI does reproduce or fully recapitulate such biological systems, i.e. whether AI approaches can be built that represent a “reverse engineering” of human intelligence.
  - We use the term **AI system for a wide variety of applications in health in line with the WHO’s document on AI ethics and governance in health** ([WHO, 2021](#)), e.g. for drug discovery, pattern recognition in health data mining etc.
  - However, for **AI-enabled software used with a medical purpose or health care purpose**, it is necessary to make specific considerations (see → [AI-enabled medical device software](#)). Guidance on medical device software – hardware combinations is available from the EU’s medical devices coordination group ([EU MDCG, 2023](#)).

#### **Definitions of ‘AI system’:**

Various definitions of AI have been proposed (for a compilation, see [European Commission, Joint Research Centre, 2020](#)), including in the EU’s high-level expert group document on ethics guidelines for trustworthy AI ([EU HLEG, 2019](#)).

We alert to the definition of “AI system” used in the European Union’s AI Act, globally the first legislation to regulate AI in a comprehensive manner:

*“AI system is a machine-based system 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.”* ([EU, 2024](#))

The definition used in the EU’s AI Act closely aligns with the definition provided by OECD:

*“An AI system is a machine-based system 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. Different AI systems vary in their levels of autonomy and adaptiveness after deployment.”* ([OECD, 2019, amended 2024; OECD 2024b](#)).

For a useful short introduction to AI, see [Scarselio, 2019](#).

#### **Term relationship:**

Related terms:

- Artificial intelligence (scientific field)
- AI typology
- AI technique
- AI-enabled medical purpose software

## AI system component / part

### **Cluster: B.1 AI and AI systems**

#### **Concept description**

AI systems are composed of the software (e.g. the machine learning model) and other physical elements, parts or components that are necessary for the intended use, such as sensors, actuators, user interfaces, hardware and software for connectivity to the internet or other devices etc.

An AI system can be in *itself a product* (e.g. a stand-alone AI system that integrates clinical data and provides predictions for clinical decision making) or it can be a *component* of another product (e.g. an

AI-enabled software within a robotic surgery suite). For more details, see the → AI-enabled medical device software).

Further, the EU's AI Act (EU, 2024a) introduces the concept of 'safety component' in conjunction of defining high-risk AI systems (Article 6, point 1). Safety component (AI Act, Article 3, definition 14) is defined as follows "*a component of a product or of an AI system which fulfils a safety function for that product or AI system, or the failure or malfunctioning of which endangers the health and safety of persons or property.*" (see also → AI safety, → NON-MALEFICENCE, → AI risk management).

### Explanatory note

A closed-loop system for diabetes management (Kovatchev, 2018; Liu et al., 2024) can feature a variety of components such as: a) AI-enabled software (possibly AI-enabled) for processing glucose-level information acquired via a transcutaneous glucose sensor for determining the amount and timepoint of delivering insulin via an injection device; b) a transcutaneous sensors for detecting interstitial glucose levels as a proxy for blood glucose levels (interstitial glucose levels have a lag time relative to blood levels of several minutes); c) an injection device for delivering the required amount of insulin at the correct time to deal with patient-specific patterns of glucose variation during the day.

Similarly, a wearable for cardiac monitoring may be composed of several components (Duncker et al., 2021): a) ECG electrodes connected to b) a smartphone associated sensor that processes waveforms recorded (e.g. using AI-enabled software) and produces warnings in case ECG characteristics are indicative of adverse cardiac events.

### Term relationship:

Related terms:

- AI system
- AI-enabled medical device software
- AI-enabled software in a medical device
- AI-enabled software as a medical device

## AI-enabled medical device software

### Cluster: B.1 AI and AI systems

#### Concept description

The term *AI-enabled medical device software* (AI-MDSW) denotes AI-type\* algorithms serving directly or in conjunction with other products a medical purpose. From a regulatory point of view AI-enabled medical device software is a variant of 'medical device software'. For relevant guidance, see for instance IMDRF, 2013; EU MDCG, 2020, EU MDCG, 2019; US FDA, 2024b; EU MDCG 2025-6. Generally, medical device software (MDSW) can be divided into three groups. Being a specific type of software, AI-enabled software can be categorised in the same manner:

- 1) **AI-enabled software in a medical device ("AI-SIMD")**, that is 'embedded' or 'part of' or a component of a medical device and used for driving, controlling or somehow influencing a medical device in its functioning, thus contributing to its medical purpose. This group can be correlated to the EU's "*MDSW with intended purpose and claimed clinical benefit related to driving or influencing a medical device for a medical purpose*" (EU MDCG, 2020).
- 2) **AI-enabled software as a medical device ("AI-SAMD")\*\***. The key distinguishing feature from an AI-SIMD is that AI-SAMD achieves its medical purpose on its own, i.e. without another (hardware) medical device. This group can be correlated to the EU's "*MDSW with independent intended purpose and claimed clinical benefit*" (EU MDCG, 2020)

|  |   |
|--|---|
|  | <p>N.B. In the context of guidance on the interplay between the MDR and the AI Act, the EU MDCG (<a href="#">EU MDCG 2025</a>) denotes <b>AI systems with a medical purpose</b> (points 1 and 2 above) as “<b>medical device artificial intelligence</b>” (MDAI), covering various types of products (see explanatory note).</p> <p>3) In addition, there is <b>AI-enabled medical device software without a medical purpose</b>, i.e. where the software on its own has no medical purpose but nevertheless supports a device with a medical purpose. This group can be correlated to the EU’s “<i>software driving or influencing the use of a medical device (with no independent intended purpose or intended claimed clinical benefit)</i>” (<a href="#">EU MDCG, 2020</a>).</p> <p><small>*) AI-type algorithm: see concept description of → AI system.</small></p> <p><small>**) WHO refers to AI-SAMD as “<i>artificial intelligence-based software as a medical device</i>” (<a href="#">WHO, 2021b</a>)</small></p> |
| <b>Explanatory note</b>  |   |
| <p><b>Definition of 'medical device'</b></p> <p>For a definition of the term medical device we refer to the definitions provided in relevant legislation, e.g. the EU's medical devices Regulation (MDR) (<a href="#">EU, 2017a</a>) and in vitro diagnostic medical devices Regulation (IVDR) (<a href="#">EU, 2017b</a>): Article 2 "Definitions" point (1) defines "medical device" for both MDR and IVDR. Definitions in other jurisdictions may differ.</p>   |   |
| <p><b>EU's concept of "medical device software" (MDSW)</b></p> <p>The EU's MDCG guidance document (MDCG 2020-1) lays out an umbrella term for all "Medical Device Software (MDSW)". MDSW is software <i>"that is intended to be used, alone or in combination, for a purpose as specified in the definition of a "medical device" in the medical devices regulation or in vitro diagnostic medical devices regulation."</i></p>  |   |
| <p><b>EU's concept of AI systems used for medical purposes</b></p> <p>AI systems used for medical purposes are, within guidance document MDCG 2025-6, called “<i>medical device artificial intelligence</i>”, MDAI (<a href="#">EU MDCG, 2025</a>), covering Annex XVI products of the EU's medical devices regulation (<a href="#">EU, 2017a</a>) as well as accessories to medical devices, in vitro diagnostic medical devices and accessories to in vitro diagnostic medical devices.</p>  |   |
| <p><b>The EU's concept of high-risk AI system (Article 6 of AI Act; EU, 2024a)</b></p> <p>Under the EU AI Act, AI systems used for medical purposes are considered high risk. Article 6 outlines that if the AI system is a safety component of a product or a product itself and covered by EU harmonisation legislation listed in Annex I of the EU AI Act <u>AND</u> the safety component or the product are required to undergo third party conformity assessment with a view of their placing on the market under Union harmonisation legislation listed in Annex I of the EU AI Act, the component or product are considered “high-risk” under the AI Act.</p> <p>The list of legislations in Annex I of the EU AI Act includes the MDR and IVDR. Thus, an AI system used as a safety component of a medical device or in vitro diagnostic medical device or an AI system constituting a medical device itself is considered of high risk under the EU's AI Act.</p> <p>Notably, both the MDR and IVDR have a dedicated risk classification approach and devices considered high risk under the AI Act may not necessarily be in the highest risk class of either the MDR or IVDR.</p> |   |
| <p><b>Term relationship:</b></p> <p>Synonyms:</p> <ul style="list-style-type: none"> <li>• Synonym of AI-SAMD: <i>Artificial intelligence-based software as a medical device (AI-SaMD)</i>; see <a href="#">WHO (2021b)</a>.</li> </ul> <p>Related terms:</p> <ul style="list-style-type: none"> <li>• <b>AI system</b></li> </ul>   |   |

- Medical device artificial intelligence, MDAI (see [EU MDCG 2025](#))

## AI-enabled software in a medical device (AI-SIMD)

### **Cluster:** B.1 AI and AI systems

See → AI-enabled medical device software

## AI-enabled software as a medical device (AI-SAMD)

### **Cluster:** B.1 AI and AI systems

See → AI-enabled medical device software

## AI system output

### **Cluster:** B.1 AI and AI systems

#### Concept description

Output of an AI system (→ [output and output data](#)) means the results generated by an AI system's algorithm as a result of the processing of → [input data](#). For most AI systems, only adequate → [input data](#) that meet the specifications of the AI system design will yield meaningful results. This is important for instance in radiological image processing and analysis, especially in situations where there may be drift in the post-deployment phase (→ [drift / shift in machine learning](#)).

In general → [output and output data](#) of an AI system are virtual (e.g. digital image manipulation, classification, a recommendation, a measurement, a forecast, new content, e.g. → [synthetic health data](#)) or physical (e.g. robotic surgery: movement of actuators).

Outputs of an AI system are data themselves, inasmuch as they can be used to study in a systematic manner input-output relationships (→ [intelligibility](#), → [interpretability](#) and [explainability](#)) and evaluate → [explanations of AI systems and their outcomes](#).

The output can also be termed “task” (e.g. [OECD, 2019b](#)) since many outputs relate to specific tasks that would typically require intelligence of a human.

#### Explanatory note

While an output generated by an AI system or AI model can be considered data in itself, the term → [output and output data](#), due to the growing prominence of → [generative AI](#), is increasingly understood as data that are created or synthesised by a generative AI system's algorithm and based on the processing of → [input data](#).

#### Term relationship:

Related terms:

- [Training data](#)
- [Post-deployment input data](#)

## AI typology

### Cluster: B.1 AI and AI systems

#### Concept description

An approach of categorising → **AI systems** depending on a variety of properties other than the computational approach underlying the AI system and with the choice of typologizing properties focused on a given context or for a given community or a specific debate.

#### Explanatory note

AI systems can be categorised for convenience in various types, which enables actor communities to discuss and exchange information on the essential hallmarks of a given system. It is important to keep in mind that these categories are to some extent arbitrary and do not reflect real-world entities, but conceptual simplifications. Useful categories include: tasks addressed, level of "autonomy" (→ **machine agency**, → **automation**), → **input data** processed, level of → **human oversight** required and/or possible, → **corrigibility**. AI typology may overlap to some extent with the dimension of tasks or → **output and output data** (e.g. [OECD, 2019b](#)).

Useful attributes for typifying AI include (list based on [Sarker, 2021a, 2022](#)):

- **visual AI** (i.e. AI processing visual/image/video data),
- **textual AI** (AI designed for analysing textual input),
- **multi-modal AI** (AI that has been trained with and can make use of various → **data modalities**).
- **analytical AI** (AI designed to identify, interpret and communicate meaningful patterns of data),
- **functional AI** (similar to analytical AI, functional AI analyses data, but rather than providing a recommendation, executes actions; see → **machine agency** and → **automation**),
- **generative AI** (AI that 'generates' or rather synthesizes new data or "content" from input data and based on the training data; see → **foundation models**; **generative AI**).
- **interactive AI** (AI designed for human computer interaction, HCI),

Obviously, AI systems may be characterised by more than one of these attributes. Concerning the **level of automation** (colloquially "autonomy") of the AI system, one can distinguish systems that are fully automated during their deployment stage (→ **machine agency** and → **automation**) or, at least, during specific periods of their deployment stage, from systems that require or depend on → **human agency**, e.g. supervision. Even systems that are fully automated require human agency at specific steps of their life cycle (e.g. for development, monitoring/performance checks, → **decommissioning**).

#### Term relationship:

Related terms:

- AI technique
- Output and output data
- Analytical AI
- Functional AI
- Interactive AI
- Textual AI
- Visual AI
- Generative AI

## AI technique

### Cluster: B.1 AI and AI systems

#### Concept description

Computational technique employed for realising the algorithmic model underlying an AI system. The AI technique greatly influences the level of intrinsic –interpretability of the AI system's functioning and its input-output relationships (see → intelligibility; → interpretability and explainability).

#### Explanatory note

The following description of 10 AI techniques is based on the proposal by Sarker et al. (2021b, 2022):

- → **Machine learning**, an umbrella term for techniques that use a → machine learning algorithm that is run on → training data to render a → machine learning model: the model is the output of the machine learning algorithm. Various learning forms can be distinguished, including:
  - *Supervised learning*, using labelled → training data. Its most common tasks are classification (i.e. predicting a label such as ‘positive’ or ‘diseased’ or ‘negative’ or ‘not-diseased’) and regression analysis (i.e. predicting quantitative values). Supervised learning uses pairs of **input data x** and associated labelled **output data y**. During supervised machine learning, the → machine learning algorithm is trained on **n input and output data pairs ( $x_n, y_n$ )**, allowing the resulting → machine learning model to make predictions for **new unseen input data  $x_u$** . Popular techniques include: K-nearest neighbours, decision trees, support vector machines, ensemble learning, random forest, support vector regression. To obtain prediction models that are indeed able to predict  $x_u$  data, usually a large amount of labelled training data is required, which is costly. So-called ‘weakly supervised learning’ accepts also noisy or imperfect data labels, data are less costly (see for instance [Karimi et al., 2020](#)).
  - *Unsupervised learning* also referred to as a ‘data driven method’. It uses only input data without labels. Its primary application is the discovery of patterns, structures or knowledge from these unlabelled data. *Clustering*, the definition of *association rules* and *anomaly detection* are commonly unsupervised tasks.
  - *Semi-supervised learning* blends supervised and unsupervised learning.
  - *Self-supervised learning* means that the data ‘supervise’ algorithmic ‘learning’.
  - → *Transfer learning* involves the use of models that were pretrained (e.g. for a classification task) with a large set of data from a different real-world problem (e.g. general image recognition) and subsequent fine tuning of that model in view of adapting it to the real-world problem at hand (e.g. medical image segmentation). This works in particular for problems that are of the same data modality: image recognition requires for instance recognition of basal visual features such as edges or orientations, a ‘skill’ which can be applied to any other image recognition task.
  - → **Deep learning** (a type of machine learning employing → artificial neural networks made up of layers of virtual ‘neurons’ which mimic essential input-output properties of natural neurons (e.g. information integration, threshold values). For an excellent review of the long history of deep learning neural networks, see [Schmidhuber, 2015](#)). Neural networks go back to the seminal work of McCulloch and Pitts concerning the ‘artificial neuron’ (see → computational theory of mind (CTM)). Neural networks are very powerful tools, especially for visual AI / computer vision ([Wang et al., 2024](#)). However, they are typically considered ‘black-box’ approaches that are not intrinsically interpretable and require explainability techniques to render their outputs explainable and hence intelligible / understandable.

- **Classical machine learning approaches** include decision trees, random forest, support vector machines, Bayesian networks, genetic algorithms and many others.
- **Data mining, knowledge discovery, advanced analytics**, including descriptive, diagnostic, predictive and prescriptive analytics
- **Rule based modelling and decision-making**: systems that store and modify knowledge that is represented as sets of rules (e.g. IF <antecedent> THEN <consequent>). Rule-based systems can be powerful decision-making tools with abilities matching those of humans. Notably rule-based modelling is typically not employing machine learning. Borah and Nath (2018) have used rule-based modelling for identifying risk factors.
- **Fuzzy logic-based approaches**, allowing approximate reasoning. Fuzzy logic, in deviation of classical logic which only allows true (degree of truth 1.0) or untrue (degree of truth 0.0) statements, allows degrees of truth or partial truth that are between true or untrue/false. Fuzzy logic is particularly successful in handling vague and uncertain data.
- **Knowledge representation, uncertainty reasoning and expert system modelling**. These techniques are concerned with how an intelligent agent's beliefs, intents and judgments can be mapped and expressed in machine-readable form for automated reasoning. The most frequently used approaches are ontology-based approaches, rule-based approaches and uncertainty and probabilistic reasoning.
- **Case-based reasoning** uses knowledge from previously successful solved problems (so-called 'cases') to work out strategies and reasoning for addressing new problems. There are applications for case-based reasoning in medicine, including breast cancer (Lamy et al, 2019).
- **Text-mining and natural language processing**
- **Visual analytics computer vision** and pattern recognition
- **Hybrid approach, searching and optimisation**

#### Term relationship:

Related terms:

- AI typology
- Machine learning algorithm
- Non-learning algorithm

## Conversational agents

#### Cluster: B.1 AI and AI systems

#### Concept description

Conversational agents (CAs), also known as chatbots or dialogue systems, are computer systems that communicate with users through natural language user interfaces involving images, written (text) or spoken language (voice).

*Adapted from Laranjo (2018) and Schachner et al. (2020).*

#### Explanatory note

Conversational agents may operate with various modalities of → input data (→ data modality) and are employed for various purposes in healthcare, including the management of mental health problems (e.g. depression, anxiety, autism) or cases of language impairment.

## Term relationship

Synonyms:

- Chatbots
- Dialogue systems

## B.2 Ethics, AI ethics, governance, management

### Ethical principles

**Cluster:** B.2 Ethics, AI ethics, governance, management

#### Concept description

Ethical principles are an element of applied ethics (→ AI ethics) and in particular deontological ethics ([Di Mattia, 2008; SEP: deontological ethics](#)). Deontology is in itself a subset of moral or normative (deontic) theories regarding possible choices that are ethically or morally<sup>1</sup> required ('shall'), permitted ('may') or, inversely, forbidden ('must not'), i.e. to responsibilities, duties and obligations that can be framed as ethical demands (or moral / normative demands). Legislative frameworks are practical applications of deontological ethics. We focus in the following discussion on ethical principles, currently the prevalent approach to tackle → AI ethics.

#### **Ethical principles and their relation to human rights**

In the tradition of the philosophy of European enlightenment, ethical principles are considered universal, i.e. uncoupled from subjective individual viewpoints. Ethical principles, at the most fundamental level, relate to the human rights and in particular human dignity<sup>2</sup> (→ DIGNITY, FREEDOM AND AUTONOMY) as an absolute and ultimate anchor for developing other ethical or 'human rights principles' that are universal (applicably to everybody), indivisible (do *all* apply), interdependent and interrelated (related to and build on each other).

#### **Moral philosophy and trust**

In his important book on moral philosophy "The ethical demand", Løgstrup ([1956](#)) posits that all interaction between human beings involves a basic trust and, importantly, that the content of this trust cannot be derived from any rule<sup>3</sup>.

While trust is not an ethical principle (it is impossible to formulate an ethical demand for trust), trust can be seen as a basic currency between people and as a fundamental basis for acceptance and adoption of technologies, services or products (see → TRUST AND TRUSTWORTHINESS). During the post-war period, nuclear technology was subject to intensive debate including its trustworthiness. Later, in parallel to new technologies, debates shifted to biomedical research, biotechnologies, genetically-modified organisms and, recently, AI.

Ethical demands can operationalise conditions for trust and, thus, responding to ethical principles will help gaining 'trustworthiness'. Thus, trust can be seen as an overarching desideratum requiring ethical demands to be addressed. These demands have typically been formulated as ethical "principles", with the bioethical principles of non-maleficence, beneficence, autonomy and justice (Beauchamp & Childress, 1979; see → bioethics) being the blue print for the approach in the field of AI today. Notably, the use of ethical principles has been criticised and this critique remains salient today (see → principlism).

#### **Bridging ethical principles and trust: trustworthiness as an overarching concept to be realised through ethical principles and other requirements**

In any case, basically all AI ethics guidelines and in particular those issued by international non-private organisations (e.g. three expert groups mandated by the EU Commission, OECD, WHO, NIST; Council of Europe) are making use of or reference to ethical principles or "value-based principles" (OECD, 2019). Principles can be seen as a useful tool for grouping moral questions and facilitating ethical and scientific debate. The EU high-level expert group of the EU Commission (2019) made trust and trustworthiness an explicit topic for → AI ethics.

#### **Footnotes:**

<sup>1)</sup> N.B. Ethics and morality are typically used interchangeably by ethicists. The ancient Greek word *ethos* originally denoted an accustomed place (namely for letting horses graze) but subsequently evolved to denote *custom*, *habit*, *manner* i.e. the full equivalent of *mores* (Latin) for customs and societally agreed habits. Consequently, moral or

ethical requirements (sometimes referred to as “imperatives”) can be denoted “moral demand”, “ethical demand” and also “normative demand”.

<sup>2)</sup> Dignity is the primary concept of human rights, being a human right itself but giving rise to the other human rights such as justice, freedom autonomy ([EU Charter of human rights, 2000](#)).

<sup>3)</sup> Løgstrup later developed an “ontological ethics” in contrast to the traditional deontic and other forms of ethics.

## Explanatory note

Given the importance of human rights for the formulation of ethical principles, specific ethical principles for AI ([→AI principles and AI ethics guidelines](#)) address fundamental values associated with **human rights** (e.g. → DIGNITY, FREEDOM AND AUTONOMY) and **justice** ([→ FAIRNESS; → RESPONSIBILITY](#))

They also include aspects that are intended to ensure that AI technologies are developed and deployed/used in a manner that

- **allows control and public and democratic discourse** and is not **harmful to society** as a whole. This relates to transparency to ensure open democratic debate ([→ TRANSPARENCY](#)) about the impacts on human rights and hence ultimately human dignity ([→ DIGNITY, FREEDOM AND AUTONOMY](#))
- is not excluding or harming **specific vulnerable groups and minorities**, i.e. relating to fairness, non-discrimination, diversity, inclusion ([→ FAIRNESS](#)) and solidarity with vulnerable groups of society including specific patient groups ([→ SOLIDARITY](#)).
- is **environmentally sustainable** ([→ SUSTAINABILITY](#)) or ‘sustainable’ in regard to societal aspects (e.g. workplace<sup>3)</sup> ([→ SOLIDARITY](#)).

Further, some bioethical concepts ([→ bioethics](#)) such [→ BENEFICENCE](#) and [→ NON-MALEFICENCE](#) have been incorporated to a varying degree into some proposals for AI principles (albeit typically under the term “well-being”, e.g. [OECD, 2019a](#), [WHO, 2021a](#)) either denoting individual well-being or that of society (e.g. [EU HLEG, 2019](#)) or both. Notably, the [European Group of Ethics \(2018\)](#) has included “mental integrity” in its proposed 9 ethical principles.

Clearly, for AI in health, the concept of beneficence is of critical importance from both a perspective of medical ethics as well as the requirement of demonstrable benefits, both in the context of clinical evaluation of health technologies (e.g. during clinical investigations and post-market surveillance) and more downstream health technology assessment (HTA) for purposes of reimbursement of health technologies by health systems.

## Footnote:

<sup>3)</sup> WHO (2021) subsumes workplace effects under the ethical term “sustainability”, whereas other guidance groups this under various “principles” such as “justice, equity and solidarity” ([EU, EGE group, 2018](#)), “Societal and environmental well-being” ([EU HLEG, 2019](#)), “Individual, social and environmental well-being” ([EU expert group, ethics by design, 2021](#)), “inclusive growth, sustainable development and well-being” ([OECD, 2019, 2024](#)). This illustrates, for only one term, the lack of alignment in regard to the formulation of ethical principles of ethics guidelines issued by international organisations

## Further reading

### **Important documents on human rights principles, conventions and charters, including in relation to biomedicine (Oviedo Convention):**

- Council of Europe (1950, entering into force 1953). European convention on human rights.  
Online: <https://www.coe.int/en/web/human-rights-convention>
- United Nations (1966) International covenant on civil and political rights. Online:  
<https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights>
- European Union (2000) Charter of fundamental rights of the European Union. Online:  
[https://www.europarl.europa.eu/charter/pdf/text\\_en.pdf](https://www.europarl.europa.eu/charter/pdf/text_en.pdf)

- Council of Europe (1997) Oviedo convention on human rights and biomedicine (ETS No 164).  
Online: <https://www.coe.int/en/web/bioethics/oviedo-convention>

**AI principles, ethical principles, ethics guidelines and recommendations issued by International organisations:**

- See references **Box 2** under → TRUST AND TRUSTWORTHINESS and **Box 8** under → AI principles and AI ethics

**Term relationship:**

Related terms:

- Bioethics
- AI ethics
- AI principles and AI ethics guidelines
- Human-centric AI
- Principlism

## Bioethics

**Cluster: B.2 Ethics, AI ethics, governance, management**

**Concept description**

Bioethics is a sub-field of applied ethics.

Bioethics is concerned with the possible consequences of biological and biomedical research and their results and applications. It includes the study of the possible benefits and harm of new methods, technologies and approaches (e.g. medical treatments) and the characterisation of circumstances, conditions and measures for avoiding harm while preserving benefits. For an overview, see Serna & Seoane (2016).

Contemporary bioethics encompasses various subfields: professional conduct and codes, research ethics, public health ethics, organisational ethics, and clinical ethics (Jankowski, 2014; Varkey, 2021). Bioethics and bioethical principles have been used for formulating principles for ethical artificial intelligence ("AI principles"; → AI ethics).

**Explanatory note**

A seminal work on bioethics is "Principles of biomedical ethics" (Beauchamp & Childress, 1979. Revised in 2001). The authors propose four foundational bioethical principles:

- **Beneficence** i.e. "do good" (for the patient), or more generally "create benefits" (see → BENEFICENCE)
- **Non-maleficence**, i.e. "do not cause harm" ( see → NON-MALEFICENCE),
- **Autonomy** by respecting preferences of patients, communities, users, society etc. (see → DIGNITY, FREEDOM AND AUTONOMY)
- **Justice**, i.e. respecting relevant values and applicable legal frameworks (see → RESPONSIBILITY).

Another key milestone in regard to the bioethical discourse is the Council of Europe's Oviedo convention on **human rights** and **biomedicine** of 1997 (Council of Europe, 1997a, b).

The approach of using ethical principles for building a broader ethical framework has been later termed "→ Principlism". This concept which has been **criticised for obfuscating the complexity of moral questions** and of **leaving a gap between high-level principles and more granular issues** of how

to address these. This critique points to problems that are still relevant and also concern AI ethical principles.

The four principles of bioethics have been adopted to various degrees and to various degrees of clarity (when naming relevant principles) for → **AI ethics** and have been complemented by other ethical principles (some newly devised, e.g. “→ **explicability**”) and/or “requirements” ([EU HLEG, 2019](#)), which in fact mix ethical principles (e.g. transparency, diversity) with technical prerequisites (e.g. “technical robustness”).

### Term relationship

Related terms:

- [AI ethics](#)
- [AI principles and AI ethics guidelines](#)

## Principlism

### **Cluster:** B.2 Ethics, AI ethics, governance, management

#### Concept description

An approach pioneered in → **bioethics** that makes use of a specific set of ethical principles to address moral questions specific to a domain (e.g. a technology) and to capture moral rules and ideals of particular relevance to that domain (e.g. biomedicine). The term ‘principlism’ was coined by Beauchamp & Rauprich ([2016](#)).

The principlism approach has been criticized for a number of reasons that remain relevant also in the context of → **AI ethics**: a) principles are not guides to action but rather “labels” for matters that may be superficially connected when dealing with a moral problem, b) there is no systematic relationship between principles and they may even conflict or overlap ([Clouser & Gert, 1990](#)).

Additional concerns are the gap between principles and actionable concepts for addressing the ethical demands stipulated by high-level ethical principles. This concern is of relevance for AI ethics.

#### Explanatory note

Since its introduction in bioethics, proposals of how to address → **AI ethics** have to an overwhelming extent made use of a **principlism-based approach**, i.e. ethical issues of AI have been formulated as and packaged into a series of “ethical principles”, resulting in a **high degree of heterogeneity** and a **certain level of potential discordance** in regard to a more detailed understanding of what these high-level principles entail in regard to moral and normative questions. Although several meta analyses on published ethical principles and guidelines support the notion of convergence ([Jobin et al., 2019](#)) around a common set of about ten principles (see references → [AI principles and AI ethics guidelines](#)), this does not necessarily reflect a common understanding of what principles encompass in terms of normative actions.

A further critique is that it is largely unclear how to translate high-level principles into practice ([Mittelstadt, 2019](#)). In addition, most ethical principles for AI do not take into consideration specific needs of individual domains of application (e.g. health, education, finance, algorithmic search etc.), but remain, in the case of AI, on the level of all possible application domains.

A notable exception is the WHO’s document on AI ethics in medicine and healthcare ([WHO, 2021](#)) which addresses health-care specific aspects under a set of six principles. These are not aligned with or mapped against earlier sets of AI principles (e.g. EU high-level expert group or those of the OECD).

### Term relationship

Related terms:

- [AI ethics](#)

- AI principles and AI ethics guidelines

## AI ethics

**Cluster:** B.2 Ethics, AI ethics, governance, management

### Concept description

We understand the term “AI ethics” as referring to the philosophical area of applied ethics in relation to AI. Various man-made technologies have been examined by applied ethics concerning their potential to cause both, harm and benefits. → **bioethics** is a prominent example of applied ethics in regard to a set of technologies in the area of life and health sciences. Importantly, AI ethics draws on principles formulated in the field of → **bioethics** and, consequentially, → **AI principles and AI ethics guidelines** incorporate or make use of bioethical principles.

While many technologies or activities in the health domain require ethical considerations (e.g. conducting clinical studies, use of gene technology for therapies etc.), artificial intelligence is a unique technology due to its potential to mimic, complement or even substitute → **human agency**.

This calls for a rigorous ethical interrogation of AI systems, i.e. their design, their development, the interests behind their development and deployment, their functioning (→ **intelligibility**) and their deployment and integration into existing processes and workflows, including clinical workflows and → **clinical pathways**.

### Explanatory note

#### **The AI ethics debate is not new:**

Notably, the debate about ethics of AI or ‘machine intelligence’ is not new and observations made decades ago still appear salient today: Wiener observed in his article on “some moral and technical consequences of automation” ([Wiener, 1960](#)): “*As machines learn they may develop unforeseen strategies at rates that baffle their programmers*”. Subsequently, Samuel ([Samuel, 1960](#)) refuted Wiener’s points, stating “*The machine is not a threat to mankind, as some people think. The machine does not possess a will, and its so-called "conclusions" are only the logical consequences of its input, as revealed by the mechanistic functioning of an inanimate assemblage of mechanical and electrical parts.*”

#### **Ethics and applied ethics:**

Ethics can be subdivided into three areas: 1) Metaethics (trying to approach the nature of ethical theory itself, trying to unravel the epistemological, semantic, psychological presuppositions of moral practice). 2) Normative ethics (the study of what makes actions right or wrong and, in particular, the definition of criteria for right and wrong). 3) Applied ethics, which is the application of normative ethical theories to concrete real-life situations in order to determine what is right or wrong. Applied ethics, including for AI, typically relies on → **ethical principles** (see → → **AI principles and AI ethics guidelines**). Various countries (see Annex 2 and → **AI principles**) and international organisations and expert groups working for these have made proposals for AI ethics, typically formulating ethical requirements through “principles” (→ **bioethics**).

#### **Ethics of AI:**

The ethics of Artificial Intelligence have been and continues to be a key concern of the research community, regulators, stakeholders and wider society in order to avoid the development, use and application of harmful AI systems that are not rooted in universal values, namely universal human rights.

While ethical considerations are relevant for many technologies, there are particularly relevant for AI due to its central tenet, i.e. to create machines that are equally or better suited to solve tasks that typically require human intelligence. Thus, AI touches or “infringes” on the most fundamental self-conception of being a human: our intelligence and autonomy or freedom to act based on intelligent insights

and reflections (→ [human agency](#)). To address this concern, various authors and organisations (including governmental, expert committees working for supranational bodies, academic institutions, civil society organisation, professional associations and NGOs) have proposed documents outlining ethical principles for consideration by various actors along the AI value chain (see for instance Jobin et al., 2019). This includes AI principles issued by countries and global regions, including EU, USA, India, Japan, Australia, Singapore, UK, Canada, China, and United Arab Emirates).

#### **AI ethics and human rights:**

The relationship between human rights and AI ethical principles and the potential use of human rights principles have been emphasized (e.g. [Saslow & Lorenz, 2019](#)). International human rights have been proposed as a framework for algorithmic accountability ([McGregor et al., 2019](#)). Nevertheless, some authors argue that the importance of human rights for AI ethics is still not sufficiently emphasized ([Jones, 2023](#)). Notably, relevant EU guidelines and the EU's horizontal legislations on AI ('AI Act') refer to the values laid down in Treaty of the European Union, the Charter of Fundamental Rights of the EU and international human rights law. ([European Commission – European Group on ethics in science and new technologies , 2018; EU HLEG, 2019, c.f. page 9](#)).

#### Further reading

##### Background on philosophical foundations (all references last accessed 2024.08.21):

- SEP - Stanford encyclopedia of philosophy entries
  - metaethics. <https://plato.stanford.edu/entries/metaethics/>
  - The principle of beneficence in applied ethics. <https://plato.stanford.edu/entries/principle-beneficence/>
- Britannica, entries
  - metaethics: <https://www.britannica.com/topic/metaethics>
  - applied ethics: <https://www.britannica.com/topic/applied-ethics>
  - normative ethics: <https://www.britannica.com/topic/normative-ethics>

#### Term relationship:

Related terms:

- Ethical principles
- Bioethics
- AI principles and AI ethics guidelines
- Ethical evaluation of AI

## AI principles and AI ethics guidelines

### **Cluster: B.2 Ethics, AI ethics, governance, management**

#### Concept description

'AI principles' or formulate high-level ethical desiderata and, to some extent, normative requirements. AI principles relate to and formulated usually as → [ethical principles](#). Sometimes (e.g. [EU HLEG, 2019](#)) they also address technical requirements (e.g. 'robustness'). Most published AI principles are aimed at the ethical development, use and implementation of AI in various fields or domains of application. The WHO has published a guidance on ethics and governance of AI for health ([WHO, 2021a](#)).

#### **A growing body of literature**

The last decade has seen a burgeoning body of publications on principles and guidelines for ethical AI from a variety of organisations: academia (e.g. [Floridi et al., 2018](#)) private companies, research bodies, public authorities and international organisations as well as their respective expert advisory committees (e.g. [European Commission, 2018; EU HLEG, 2019; OECD, 2019a; WHO, 2021; United Nations - Unesco, 2021; Council of Europe, 2024](#)). AI principles have been published by various countries and global regions, including EU, USA, India, Japan, China, UK, Canada, United Arab Emirates, Singapore, Australia.

Published AI ethics documents range from an examination of applicable ethical and, in some cases, bioethical principles to statements of intent (“desiderata”) or requirements regarding → AI ethics. Some of the public guidance documents have been considered “soft law” (e.g. EU HLEG, 2019; 2020) which has been related also to policy formulation of the EU’s AI Act (Laux et al., 2024). Recently, the FUTURE-AI consortium has published a consensus guideline for trustworthy AI, using six guiding principles as a framework for defining practical steps that implement the guidance recommendations (FUTURE-AI consortium, 2025).

However, while there is an apparent convergence (see below) towards basic ethical principles, there is no consistent presentation of AI ethical principles among various organisations: there is currently no consensus concerning a common set of “horizontal” AI ethical principles, aimed at a variety of possible application fields or domains. Neither is there consensus for AI principles in specific application domains (e.g. health).

#### ***AI principles as a topic of academic study:***

The resulting complexity and fragmentation of the landscape of AI (ethical) principles and ethics guidelines has become subject of academic study.: Several systematic reviews () have examined this growing body of literature, identifying the most frequently used ethical principles, examining to which extent these are used in practice and identifying obstacles in regard to the practical implementation of → AI ethics (Jobin et al., 2019; Hagendorff et al., 2020; Fjeld, 2020; Ryan & Stahl, 2021; Kluge Corrêa et al., 2023). In addition, there are a number of projects aiming at collecting documents on ethical principles and/or cross-mapping or link ethical principles for AI (e.g. SEA platform network, 2024).

#### ***Inconsistent presentation of AI principles and the interpretation gap:***

While the various publication have provided a much-needed basis for discussion and, in specific cases, provided even practical guidance (e.g. in form of checklists, e.g. EU HLEG, 2020; see also Radclyffe, 2023), the *fragmentation* of AI principles is not conducive to a consistent use and application of → AI ethics and more detailed and much-needed discussions on how to translate ethical principles into practice are hampered by this lack of a common conceptual framework, which is prone to create confusion among key stakeholders and delays in regard to the development of pragmatic and actionable tools for implementing → AI ethics in the daily practice of AI development, deployment and use. Further, the apparent convergence may hide *interpretative differences* of conceptual and normative nature (Mittelstadt, 2019; Widjaja, 2024; see also Whittlestone et al., 2019). There is thus a risk that, given this interpretation gap, the development of guidance or standards may yield different practical results or that efforts are slowed down by an insufficient conceptual foundation of AI ethics with negative effects on the development of pragmatic approaches towards the overarching aim of trustworthy AI.

#### **Explanatory note**

For an overview of ethical or value-based AI principles published by public bodies or their expert panels, see → TRUST AND TRUSTWORTHINESS.

#### ***Platforms collecting and linking ethical principles (non-exhaustive):***

- The **SEA Platform network provides a useful website aiming to link the various AI principles** (“Linking AI Principles”, LAIP). Online: <https://www.linkinai-principles.org/>
- **Algorithmwatch.org** provides an inventory of various documents, the “AI Ethics guidelines global inventory” (e.g. governmental, private sector, supranational) that cover or touch on ethics for AI. Online: <https://inventory.algorithmwatch.org/>
- The **Council of Europe’s website on “AI initiatives”** lists about 450 documents 450 documents related to artificial intelligence, coming from national authorities, the private sector, international organisations or multi-stakeholder initiatives. Online <https://www.coe.int/en/web/artificial-intelligence/national-initiatives>. The site allows accessing a non-exhaustive data collection on AI documents. Online: [https://docs.google.com/spreadsheets/d/1mU2brATV\\_fgd5MRGfTASOFepAI1pivwhGm0VCT22\\_U/edit?gid=0#gid=0](https://docs.google.com/spreadsheets/d/1mU2brATV_fgd5MRGfTASOFepAI1pivwhGm0VCT22_U/edit?gid=0#gid=0)

- **The OECD AI policy observatory provides an overview of policies, data and analysis for trustworthy AI.** Notably, it also provides various proposed **tools and metrics for trustworthy AI**. Online: <https://oecd.ai/en/>

### Term relationship:

Related terms:

- AI ethics
- Ethical principles
- Bioethics

### **Box 8.** AI principles, ethical principles, ethics guidelines and recommendations issued by international organisations

#### **United Nations**

- UNESCO (2021) Recommendation on the ethics of artificial intelligence. Online: <https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence>
- UNESCO (2024) Ethics of Artificial Intelligence. Online: <https://www.unesco.org/en/artificial-intelligence/recommendation-ethics>

#### **EU**

- European Commission – European Group on Ethics in Science and New Technologies (2018) Statement on artificial intelligence, robotics and ‘autonomous’ systems. Online: <https://op.europa.eu/en/publication-detail/-/publication/dfebe62e-4ce9-11e8-be1d-01aa75ed71a1> (Last accessed: 2024.08.21)
- EC HLEG (2019) European Commission high-level expert group on artificial intelligence: Ethics guideline for trustworthy AI. Online: [https://ec.europa.eu/futurum/en/ai-alliance-consultation\\_1.html](https://ec.europa.eu/futurum/en/ai-alliance-consultation_1.html)
- European Commission (2021): Ethics by design and ethics of use approaches for artificial intelligence Online: [https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence\\_he\\_en.pdf](https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence_he_en.pdf)
- EU (2024) Regulation laying down harmonised rules on artificial intelligence (“AI Act”). Online: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689>. For a version that can be easily navigated, see <https://artificialintelligenceact.eu/the-act/>

#### **World Health Organisation (WHO)**

- World Health Organisation (2021) Ethics and governance of artificial intelligence for health. Online: <https://www.who.int/publications/i/item/9789240029200>
- World Health Organisation (2024) Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models. <https://www.who.int/publications/i/item/9789240084759>

#### **OECD**

- Organisation for economic cooperation and development, OECD (2023) Recommendation of the Council on Artificial Intelligence, 2023, OECD/LEGAL/0449 (2019). <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>
- Organisation for economic cooperation and development, OECD (2024) AI in health: huge potential, huge risks. Online: [https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/01/ai-in-health-huge-potential-huge-risks\\_ff823a24/2f709270-en.pdf](https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/01/ai-in-health-huge-potential-huge-risks_ff823a24/2f709270-en.pdf)

#### **„FUTURE-AI“**

- FUTURE-AI consortium (2025): international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ. doi: 10.1136/bmj.r340. Erratum for: BMJ. 2025 Feb 5;388:e081554. doi: 10.1136/bmj-2024-081554.

## Ethical evaluation of AI

**Cluster:** B.2 Ethics, AI ethics, governance, management

### Concept description

#### ***Ethical evaluation of AI in relation to AI management and AI governance***

AI systems are the first technology that reproduce or mimic → **human agency** and, consequentially, there are numerous ethical implications in relation to developing, deploying and using AI systems (Mittelstadt et al., 2016; Floridi et al. 2018; Morley et al., 2020b)

Ethical evaluation of AI means the consistent assessment of an AI system against → **ethical principles** or 'AI principles' in order to ensure that the system's development, deployment and use is aligned with these principles and their underpinning values. This should be done by the relevant community developing, deploying and using an AI system. It requires common conceptual understanding and agreement on pragmatic approaches. Ethical evaluations of AI system development require a **proactive and collaborative approach** of the relevant community along the → **life cycle of AI** in health and across the → **value chain of AI** (see also → **AI evidence pathway for health**).

Importantly, an ethical evaluation will inform processes of → **AI risk management**, since many AI-associated → **risks** relate to ethical aspects: for instance, AI systems exhibiting discriminatory bias (→ **FAIRNESS**) can pose a safety risk (→ **AI safety**) if an AI system is less accurate for a specific patient group than for another. It may also inform other relevant impact assessments (see below).

We consider ethical evaluation of AI part of → **AI management** and → **AI governance**. We propose the → **AI evidence pathway for health** as a conceptual framework for ethical evaluation of AI and for identifying the required evidence that needs to be generated and/or collected. The evidence needs will vary depending on the specifics of the AI system (e.g. for research purposes versus healthcare).

Ethical evaluations of AI relate to other concepts of AI assessment:

- → **AI impact assessment (AI-IA)**
- → **fundamental rights and algorithm impact assessment**
- **Audits** (→ **auditability and auditing**)
- **Data protection impact assessments** (→ **AI impact assessment (AI-IA)**)

#### ***Ethical evaluation is not only about technologies but also about actors and organisations***

Interrogating and evaluating an AI system in terms of ethical implications requires an understanding of the many decisions (taken by a network of human actors) that led to the transition from conception to final 'product' (→ **algorithm-to-model transition**).

Developers, → **machine learning engineers** and **data scientists** (→ **AI practitioners**) greatly influence an AI system's model, e.g. through building the → **development data** (identifying data, data wrangling), choosing → **attributes**, → **proxies** (e.g. to simplify model development) or → **features** ("feature engineering"), defining model parameters, deciding the → **model calibration**, choosing → **evaluation metrics** or other suitable methods (e.g. predictive power or selected benchmarks). Other elements are the use of value chain enablers or assets, with implications for distributed responsibilities.

Any of these decisions can have implications on ethical aspects, notably related to

- **FAIRNESS** (- > **bias**),
- **PRIVACY PROTECTION** (both of general personal information and sensitive health data),
- **NON-MALEFICENCE** (see also → **risks**, → **AI safety**),
- **RESPONSIBILITY** (e.g. clarity of → **accountability structures**),
- **TRANSPARENCY** and technical aspects, such as robustness or interoperability (see → **NON-MALEFICENCE**).

Importantly, also other actors may influence the shaping of an AI system: users (e.g. through → **usability considerations**), stakeholders (through expectations and feedback). AI systems themselves (e.g.

through automated programming, data identification, data processing, data labelling etc.) may contribute to the → algorithm-to-model transition, with possible impacts on ethical aspects.

Importantly, the way AI systems are used (e.g. in healthcare settings or health systems, may have ethical implications, independent of the technology per se.

Ethical evaluations are particularly critical for AI systems employed for consequential decision-making, e.g. in clinical (high-risk contexts) and where human dignity, freedom of choice according to e.g. a patient's preferences and → human agency (e.g. of healthcare professionals) are at stake (see → DIGNITY, FREEDOM AND AUTONOMY).

### Explanatory note

The extent of an ethical evaluation of an AI system may vary depending on its purpose and use context. Whenever AI might affect people directly or indirectly, an ethical evaluation should be conducted xxx, also to build trust. This includes all applications in healthcare and clinical practice. Clinical ethics can serve as a useful foundation for considerations of AI in health.

In contrast, AI systems that are applied for processes that do not affect people may require an ethical evaluation that is tailored to the given situation. Examples in the health domain are AI systems used to monitor inventories, stocks or to manage procurement of medical equipment and medicinal products. For such applications specific ethical principles are likely not applicable or need to be evaluated in the given context (e.g. (data) privacy, non-discrimination etc.). We recommended assessing as early as possible during the development process to which extent an AI system requires an ethical evaluation.

In the present context, a reflection on types of ethics issues is useful. Morley et al. (2022) identify three basic issues in their review of the ethics of AI in healthcare: **(1) Epistemic issues**, i.e. issues of AI that are related to misguided, inconclusive, inaccurate or inscrutable evidence that seriously affect accountability, intelligibility and, most importantly, the real-world relevance of an AI system. **(2) Normative** issues related to unfair outcomes (e.g. discriminatory, biased, non-inclusive etc.) or to transformative effects (e.g. algorithmic outputs that re-conceptualise real-world facts in unexpected ways). **(3) Traceability** issues, e.g. issues where adverse outcomes or harm are difficult to detect and trace their root causes. This may have serious consequences on identifying accountable actors (e.g. developers versus users) seriously impair contestability (challenging outcomes) and even liability.

**Traceability** issues are closely linked to the level of intelligibility of outcomes (i.e. why the system has produced a specific outcome (e.g. positive prediction) as opposed to an alternative outcome (e.g. negative prediction). Intelligibility relates to interpretability or explainability. Insufficient understanding of input-output relationships of an AI system undermine traceability, even in cases where important traceability elements such as logging and record-keeping have been observed. It should be noted though that the level of satisfactory explanations may vary depending on the context of the AI system. In healthcare and medical practice, for instance, fully mechanistic explanations may not always be available (c.f. Herzog 2022; London 2019) and explanations may have to rely on circumstantial plausibility reasoning and past evidence as well as robust experience.

### Term relationship:

Related terms:

- AI ethics
- AI principles and AI ethics guidelines
- AI impact assessment (AI-IA)
- Fundamental rights and algorithmic impact assessment (FRAIA)

## AI impact assessment (AI-IA)

**Cluster:** B.2 Ethics, AI ethics, governance, management

### Concept description

#### **Algorithmic impact assessments and related concepts**

AI impact assessments (AI-IAs) frameworks are tools for assessing AIs with the aim to identify both positive and negative impacts (benefits and harms / risks) on individuals, groups, communities, society and the environment. AI-IA relates to

- the concept of → ethical evaluations of AI, since risks cannot be disentangled from ethical aspects and risks related to ethical principles (see → NON-MALEFIENCE for risk types) and
- to → fundamental rights and algorithm impacts assessments.
- audits (for comparison, see [Sloan & Moss, 2023](#)). See also → auditability and auditing.

Currently there is no full consensus regarding a sharp delineation ethical evaluation, algorithmic impact assessment and fundamental rights impact assessments ([Stahl et al., 2023](#)). An analysis of communalities and differences might help determining one single approach covering various aspects of these assessments. The field of impact assessments is subject to ongoing debate and proposals, including on metrics ([Jenkins & Nericcio, 2023](#)). For a general orientation on AI-IAs, see [Selbst, 2021](#), [IAIA, 2024](#); [Stahl et al., 2023](#); [Institute for the future of work, 2024](#).

#### **AI-IA and non-maleficence**

AI-IAs can be considered part of a → AI governance system, reflecting in particular the ethical principle of → NON-MALEFIENCE. In regard to AI in health and depending on the situation, AI-IAs may be conducted for specific AI systems throughout the life cycle ([Ada Lovelace Institute, 2022](#)). Alternatively, they may be conducted for specific types or classes of (high-risk) AI systems (e.g. multi-modal models, → foundation models and → generative AI) that may be used in health systems.

#### **Legal frameworks and impact assessments**

There may be legal provisions for specific impact assessments. The EU's AI Act requires, for high-risk AI systems, fundamental rights impact assessments under Article 27 ([EU, 2024a](#); [Busch et al., 2024](#)) that covers various aspects, including identification of possible harms for persons or groups, human oversight (→ ensuring human agency and oversight) and harm/risk mitigation measures. Other forms of impact assessment may be related to the use of AI: For instance, the EU General Data Protection Regulation (GDPR; [EU, 2016](#)) stipulates requirements for a "data protection impact assessment" under specific circumstances (Article 35), in particular where (fundamental) rights and freedoms are concerned ([Kaminski et al., 2020](#)). This applies also to data used in the context of AI systems.

#### **Contribution of AI-IAs to areas of concern**

Apart from informing about various impacts on people, groups, communities and society, AI-IAs can be useful from various perspectives:

- **Transparency:** AI-IAs can enhance → TRANSPARENCY by providing robust foundations for the → disclosure of how organisations develop (e.g. businesses) and also use AI systems (e.g. healthcare settings and health systems).
- **Risks:** AI-IAs can be useful tools for underpinning and informing risk evaluations (including those concerning data protection and privacy), e.g. in the context of → AI risk management. AI-IA may help asking the right questions that may elude standardized and more rigid risk management procedures.
- **Benefit-risk assessments:** AI-IA may contribute to benefit-risk assessments which support determining acceptability of risks of healthcare products (see → AI risk management). AI-IA may support adjustments towards minimization of negative impacts (harms) and the

- maximization of (clinical) benefits of an AI system, including → use context and → use environment considerations.
- **Agency, oversight, patient primacy:** AI-IA may contribute to detecting lack of sufficient → human agency and → human oversight (see also → ensuring human agency and oversight) as well as problems in regard to → DIGNITY, FREEDOM AND AUTONOMY, including, in healthcare, → patient primacy.

#### Explanatory note

Stahl et al (2023) have identified 38 (as of May 2024) impact assessment frameworks that are targeted on AI, with other frameworks being either "relevant to AI" or being "possibly relevant for AI". Examples of frameworks with possible relevance (for further references see Stahl et al., 2023) include

- human rights impact assessment ([Lindblad & Bloch Veiberg 2020](#)) or
- "fundamental rights and algorithm impact assessment", FRAIA ([Dutch government, 2022](#)).
- Frameworks focused on privacy impact assessments ([Clarke 2009](#); [Wright and Friedewald 2013](#))
- technology ethics impact assessment ([Wright 2011](#)).

An important finding of [Stahl et al](#) is the notion that, currently, there is no consensus on content, structure and implementation of AI-IAs. Nevertheless, AI-IAs are important tools for stimulating reflection and discussion concerning the societal and ethical consequences of AI.

#### Term relationship:

Related terms:

- Ethical evaluation of AI
- AI principles and AI ethics guidelines
- Fundamental rights and algorithmic impact assessment (FRAIA)
- Auditability and auditing

## Fundamental rights and algorithm impact assessments

#### **Cluster:** B.2 Ethics, AI ethics, governance, management

#### Concept description

Depending on specifications, potential impact (e.g. multiplier effects due to widespread use) and risks posed, AI-based as well as non-AI based algorithms may require an assessment of their possible impacts on fundamental human rights. In contrast to → ethical evaluation of AI which encompass several → AI principles and AI ethics guidelines (including also those derived from → bioethics, e.g. → beneficence, → non-maleficence), such assessments focus on the identification of impacts on fundamental human rights.

Fundamental rights and algorithm impact assessments relate to other concepts, notably

- The concept of → AI impact assessment (AI-IA)
- Audits (see → auditability and auditing).

#### Explanatory note

The **EU AI Act Article 27** requires that, for high-risk AI systems, specific deployers and private entities providing public services must perform, prior to deployment, an assessment of the possible impacts of an AI system on fundamental rights ([EU, 2024a](#)).

The **Dutch government** has proposed a framework of “fundamental rights and algorithm impact assessment” (FRAIA). It can be used at early stages of designing an algorithm or AI system but equally for algorithms already laid out. It poses questions that should be discussed, ideally, by multidisciplinary teams “*in any instance where a government organisation considers developing, delegating the development of, buying, adjusting and/or using an algorithm.*” ([Dutch government, 2022](#)).

The questions posed in this document could also support the ex-ante assessment of potential impacts on fundamental rights resulting from the use of AI systems in various settings.

### Term relationship:

Related terms:

- Ethical evaluation of AI
- AI principles and AI ethics guidelines
- AI impact assessment (AI-IA)
- Auditability and auditing

## Alignment

### Cluster: B.2 Ethics, AI ethics, governance, management

#### Concept description

Alignment can be seen as an approach for addressing the considerable potential social and ethical risks of AI. While also ‘traditional’ AI that provides predictions in regard to real-world objectives and problems needs to be aligned to societal desiderata (e.g. non-discrimination), alignment has gained particular traction in regard to → generative AI that has been trained on data distributions and predicts novel data points (→ see also → foundation models).

Alignment refers to the *ex-ante* and continuous process of ‘aligning’ AI to human values or preferences and social desiderata ([Lipton, 2018](#)). Aligning AI is intended to enhance the overall utility of the AI for the user and address biases, “toxicity” (e.g. wrong or malevolent generative AI outputs) and privacy issues associated with the vast number of training data ([Shen et al., 2023](#)).

Consequently, alignment aims at aspects such as truthfulness, ‘interestingness’, harmlessness and safety. Alignment aims to address issues with the content synthesized (‘generated’), e.g. *undesirable content* and *unfaithful content*.

Several studies that have demonstrated that large-language models (based on issues with the training data) produce undesirable content, namely stereotypes and bias related to gender, culture and ethnicity/race (e.g. [Nadeem et al., 2020](#)). So-called ‘**unfaithful content**’ relates to content outputs that are not relating to truthful facts or events but are fabrications (‘hallucinations’ or ‘confabulations’) which may propagate mis- and disinformation. If not properly controlled, this LLM feature can be exploited by malicious users for spreading disinformation and ‘fake news’: LLMs unfortunately may provide easily scalable and cost-effective means for targeted disinformation campaigns, e.g. for political or commercial motives.

→ **When considering hallucinations, it is important to consider that all output of generative AI is a prediction or in a certain sense a confabulation or hallucination. Ideally, predictions based on the huge set of training data should have contextual meaning that aligns with facts and knowledge encapsulated in the data. In some cases however, predictions of the model do not align with facts (e.g. “predicted” publications that do not exist in reality). This is an inherent downside of generative AI that will be, likely, impossible to completely control. Knowing how to use**

**generative AI responsibly is key for mitigating potential risks from data predictions retrieved from such models.**

#### Explanatory note

Technically, alignment can be addressed via various methods. The most common one is reinforcement learning from human feedback (RLHF), involving the fine tuning based on the perceived relative quality of the model's output ([Ziegler et al., 2019](#)). Another approach is based on 'representation engineering', involving the identification of meaningful representations of human preferences that are embedded in the patterns of activity (based on their actual use) within a large language model. These representations can then be adjusted to control the outputs of the model ([Liu et al., 2024](#)).

#### Term relationship:

Related terms:

- AI ethics
- AI principles and AI ethics guidelines
- Ethical evaluation of AI
- AI impact assessment (AI-IA)
- Fundamental rights and algorithm impact assessments

## Ethics code

**Cluster:** B.2 Ethics, AI ethics, governance, management

#### Concept description

An AI ethics code is an internal or published policy document or statement outlining its approach to, as relevant, the development, deployment and use of AI systems (see also → disclosure).

An ethics code can for instance document the **intention** of an organisation and may refer to **relevant fundamental values or ethical principles, applicable frameworks** (e.g. → AI governance, → AI management) and how these are addressed.

Further, an ethics code may refer to **professional codes of ethics**, e.g. of medical associations (e.g. [Geis et al., 2019 for AI and radiology](#); [American Medical Association, 2024](#)). In contrast to a guideline, which outlines how to achieve ethical AI, a code can be seen as a statement of intent and a self-declaration of awareness of ethical aspects and frameworks used to realise ethical and trustworthy AI.

[Madary & Metzinger \(2016\)](#) have made recommendations for a code of ethical conduct for good science and use of virtual reality (VR) technology. These are relevant also for VR use in medicine and healthcare.

#### Explanatory note

As noted by the Commission high-level expert group, an ethics code cannot replace ethical reasoning to be conducted by organisations involved in developing and deploying AI systems.

#### Term relationship:

Related terms:

- AI principles and AI ethics guidelines
- AI ethics
- Ethical principles
- AI governance

- AI management

Synonyms:

- AI value platform
- AI principles  
(see for instance the AI principles of Google Inc. Online:  
<https://ai.google/responsibility/principles/>)

## AI governance

**Cluster:** B.2 Ethics, AI ethics, governance, management

### Concept description

#### **AI governance: an emerging concept**

Governance generally refers to the manner in which something is controlled, including the tools, systems, processes or frameworks used to execute such control. The term AI ‘governance’ has, depending on context, slightly different notions. It can refer to explicit and implicit *practices* for controlling AI development and/or use within private organisations / industry (e.g. [IBM, 2024](#)) or to the *overall framework* of applicable legislations, frameworks, policies and practices that guide the development and implementation of AI to ensure its responsible and ethical use (e.g. [Wagner, 2024](#)). Ethical, legal and technical opportunities and challenges have been mapped by [Cath \(2018\)](#).

Moreover, there is significant variation in regard to the *scope* of AI governance. Depending on organisation, AI governance can include notions of ethics, safety, respect for human rights and societal values during development and application/deployment of AI tools ([IBM, 2024](#)), include other ethical principles, such as sustainability ([World Economic Forum, 2024](#)) or focus on specific ethical aspects, e.g. responsibility ([TechTarget, 2024](#)). The United Nations have published an analysis of the UN system’s institutional models, functions, and existing international normative frameworks applicable to AI governance, which touches on a variety of salient points relevant for the emerging concept of AI governance ([UN, 2024](#)).

#### **Definition of AI governance and its relation to the AI evidence pathway for health**

- **AI governance** is understood here as all principles, policies, legislations, guidance, frameworks and standards used by a private or public organisation and communities for ensuring ethical and hence trustworthy AI development and use during all applicable stages of the life cycle stages and with regard to relevant value chain elements that are required for the development and use of AI tools. A recent mapping of AI governance in health has examined, amongst, other aspects on global AI governance interoperability in health ([HealthAI, 2024](#)), underlining the need for aligning AI governance also in the health domain (see also [Gill, 2021](#)). Given the fact that frameworks, guidance and standards still need to be developed and given the fact of rapid change of the science and practice of AI (→ AI as a scientific field), AI governance is and will likely remain a ‘moving target’.
- **AI evidence pathway for health:** We propose the → AI evidence pathway for health as a framework for identifying and generating necessary evidence for trustworthy AI, thus feeding into AI governance in the health area. The pathway addresses also three additional aspects (encapsulated in translational concepts of ethical principles): i) respect for democratic values and debate ([OECD, 2019/2024](#)), ii) the public interest ([WHO, 2021a](#)) and iii) safe innovation ([Council of Europe, 2024; OECD, 2019/2024; EU, 2024](#)).

- As laid out in this ontology, **ethical and trustworthy AI in health can be based on nine fundamental and granular consensus ethical principles** (ontology Part A): beneficence, non-maleficence (and thus safety), dignity (and thus fundamental human rights), privacy, transparency, fairness, responsibility, solidarity and sustainability.
- For inscrutable and complex AI tools, so-called ‘ethical auditing’ has been proposed ([Cath., 2018; Mittelstadt, 2021](#)) as a governance tool. Effective ethical auditing will however depend to a large extent on the level of understanding, documentation and knowledge about AI systems, i.e. their functioning and how and why they produce certain outputs. It thus requires approaches to enhance → **intelligibility of AI systems**, mainly through overall → **TRANSPARENCY**, as well as → **interpretability and explainability**).

AI governance requires → **AI risk management** and, for health, also a structured approach for creating evidence on benefits (→ **BENEFICENCE**), since the acceptability of risks of health products (e.g. AI-enabled medical devices software) is typically judged against the benefits of the product. We consider AI risk management risks as part of → **NON-MALEFICENCE**.

Within organisations, governance “principles” may spell out how an organisation or community tackles various aspects of governing its AI life cycle. AI governance can be implemented through the establishment of formal → **AI management processes**.

### **Public, overarching frameworks and corporate frameworks**

The OECD AI policy observatory ([OECD, 2021a](#)), maintained jointly with the European Commission, provides a continuously updated dashboard of AI governance initiatives structured in four segments: 1) governance, 2) guidance and regulation, 3) financial support and 4) AI enables and other incentives. As of January 2025, the dashboard listed about 1184 documents on guidance and regulation as well as AI governance. The dashboard also counts national legislative initiatives, including of EU member States (e.g. Germany’s “automated vehicles bill” in the road traffic act, Netherlands “experimental law on self-driving vehicles”). The dashboard illustrates the considerable dynamic of global AI governance developments. For references to *analyses* of private and public guidelines on AI principles (an important part of governance), see → **AI principles and AI ethics guidelines**. The dashboard illustrates the need to harmonise AI governance internationally, which would facilitate acceptance of AI-based products, including social media platforms utilising AI, AI-powered search engines and recommender systems, large language models etc.

### **Published legislations or executive acts**

- The EU Act ([EU, 2024a](#)) has introduced a comprehensive legislative governance structure for AI based on a risk-based approach, stipulating risk-related requirements and referring, where necessary, to existing EU legislations in regard to AI-enabled products of a specific sector (e.g. medical devices and in vitro diagnostic medical devices), providing clarity and certainty to industry and stakeholders on essential high-level requirements. Notably, the AI Act does not and cannot regulate all aspects of AI, notably the way AI is being implemented in clinical environments and workflows. Thus, additional effort is needed to introduce guardrails for the responsible use of AI in health ([OECD, 2024d](#)). For comments on the AI Act in regard to the risk-based approach, see [Fraser et al., 2023](#) and [Ebers, 2024b](#).
- South Korea has, as the second AI legislation globally, in December 2024 passed the Act on the Development of Artificial Intelligence and Establishment of Trust (AI Basic Act) ([Shin & Kim, 2024; ChosunBiz, 2024](#)).
- Although formally not a legislation, the US executive order 14110 ([White House, 2023](#)) might have resulted in a comprehensive federal governance framework for AI within the USA ([Coglianese, 2024](#)). It was rescinded in January 2025 ([White House, 2025](#)). Due to the lack of a federal legislation, several US States have already introduced AI-related legislation ([BCLP LLP, 2024](#)). This trend towards ‘technology federalism’ might now accelerate ([Kohler S, 2025](#)), leading to potential governance uncertainties within the US market ([Hyzy, 2024](#)). While a globally harmonised approach to AI legislations might address uncertainties of industry and

stakeholders and avoid a diktat of the most powerful private actors (Smith C, 2025), AI governance should not rely only on legislation, but also on collaboration between communities and clarification of existing guidance in specific sectors (e.g. medical devices, education etc.).

These public overarching frameworks will be mirrored over time in internal governance frameworks within companies and organisation to ensure regard to legal requirements while enabling implementation with a focus to the specifics of the company's AI product or AI use. Some of these internal frameworks will be adaptations of previously issued industry 'governance' documents and principles.

### Explanatory note

While legislative frameworks provide a general framework for AI governance by stipulating high-level requirements for AI systems and in particular those posing high risks, many details will need to be elaborated, e.g. through 'soft law', such as regulatory, technical and scientific guidance, good practice approaches, standards or 'common specifications' (EU Act, Article 41; EU 2024a; 2024b).

Important examples are the guidance documents of various expert groups of the European Commission (see section A.0 Trust and trustworthiness), the US executive order ([White House, 2023](#); rescinded in 2025; comment by [Coglianese, 2024](#)), US NIST's AI risk management framework ([National Institute of Standards and Technology, NIST, 2023](#)) or China's AI safety governance framework ([National Information Security Standardization Technical Committee \(TC260\), 2024](#)). While not legally binding, these documents influence concepts, priorities and practices and have already influenced (e.g. the EU's AI Act) or may influence future legislative initiatives.

In principle governance of AI is subject to various pressures (e.g. stakeholders, market pressures, legislative and regulatory). [Latzer et al. \(2014\)](#) have proposed three main angles of governance in relation to algorithms:

- a) Market based (e.g. pressures from value-chain partners, consumers; for health care this would include reimbursement for health technologies).
- b) Self-organisation (through principles, internal documents, guidance etc.) within organisations or communities.
- c) (Co-)regulation, i.e. through normative requirements set in legislations and regulatory guidance that elaborates on practical aspects of legislations.

In reality, AI governance will always involve a mix of the three angles. However, depending on global area, there may be a different weighting of these approaches.

Clearly, there are challenges to all three approaches. For instance, monopoly situations may weaken market-based approaches. Self-organisation may be opaque and inconsistent across a specific sector. Documents outlining "AI management" tools (e.g. standards) are not directly accessible to everybody (e.g. behind paywalls) and may not satisfy scientific principles and (epistemic) traceability (e.g. lack of references, unclear authorship) as required by the advancement of regulatory science.

There are also numerous challenges for (co-)regulation, including inherent uncertainty associated with definitions of AI and, linked to this, the fast pace of technological change (see for instance [Governance.ai, 2014; S&P, 2024](#)) which challenges the ability to keep up with legislative approaches and associated regulatory science and regulatory practices.

### Term relationship:

Related terms:

- [AI management](#)
- [AI risk management](#)
- [AI safety](#)
- [AI ethics](#)
- [AI principles](#)

## AI management

**Cluster:** B.2 Ethics, AI ethics, governance, management

### Concept description

AI management refers to processes of putting → AI governance into practice. Thus, while → AI governance focuses on the collection of applicable concepts, AI management focuses on the practical processes, procedures and practices to enable effective governance.

AI management systems will typically be set up in relation to defined life cycle stages (→ life cycle of AI in health), from development over deployment and post-deployment activities in view of monitoring and managing all relevant AI aspects, including → decommissioning / retirement. Key components of an AI management system include:

- → AI risk management to identify and control various risks, i.e. risks for persons; risks for organisations, groups, communities and society as a whole; risks for property and assets; risks for the environment. → AI risk management supports → AI safety (i.e. safety to persons, including patients, users and healthcare professionals).
- AI safety of patients, users and other persons relates to the broader ethical obligation of NON-MALEFICENCE. → Risks can be associated with ethical principles 3 to 9 (see → NON-MALEFICENCE) in order to generate evidence for benefit-risk ratio assessments and risk estimations.
- → Data governance
- Fulfilling internal and external documentation and transparency needs, e.g. internal → traceability for quality assurance, product improvement and for audits; legally required documentation needs including to deployers and users (including → failure transparency).

Given the highly distributed responsibilities that come with AI, AI management processes need to consider also the → value chain of AI and its agents and communities (e.g. data, IT infrastructure, enabling technologies, (pretrained) models, AI systems, processes and services required for the operation of AI systems (see. → RESPONSIBILITY).

This will ensure attributing responsibilities ([CERNA, 2018](#)) and enable the collection or generation of relevant evidence, e.g. on the provenance of data, potential bias assessment in data collections that were purchased from a provider or the potential biases of pretrained purchased AI models that are subsequently trained through → transfer learning.

### Explanatory note

→ AI governance can be achieved through a formal AI management ‘system’, i.e. a collection of approaches, processes and procedures within a company or organisation or without such formal system, e.g. through following relevant guidance documents on AI governance. In the latter case, organisations should nevertheless keep records concerning the guidance they use for managing their AI development and/or use.

[ISO \(2023b\)](#) has published a standard on an AI management system.

### Term relationship:

Related terms:

- AI governance

## AI risk management

**Cluster:** B.2 Ethics, AI ethics, governance, management

### Concept description

#### **AI risk management**

With AI risk management we mean all risk management activities that aim of identifying, framing and controlling or mitigating risks associated with AI. However, such activities will typically be part of more general risk management processes, encompassing also non-AI aspects of a product or process. AI risk management can be considered part of → AI management.

AI risk management should cover the entire life cycle (→ life cycle of AI in health) and consider also value elements (→ value chain). Additional approaches may be required to control risks related to the implementation of AI systems in specific clinical workflows, → use environments and → use contexts.

#### **AI risk management in the legal context**

The establishment of a system for risk management is a legal requirement under Article 9 of the EU's AI Act ([EU, 2024a](#)). This requirement needs to be considered in connection to existing relevant requirements of other EU legislation applicable in the context of AI in health, such as the medical devices (MDR) and in vitro diagnostic medical devices (IVDR) regulations of the EU ([see for instance MDR Annex I, Chapter I, point 3, EU, 2017a](#)).

#### **Risk management: general concept**

Risk management refers to all activities aimed at identifying, framing, controlling and mitigating risks – in this context in relation to AI systems. While a central purpose of risk management is to ensure safety of patients, users and other persons (→ AI safety; → NON-MALEFICENCE), risk management generally has a broader scope. It covers variety of risks: safety risks, but also risks for organisations, groups, communities, society as a whole, property and the environment.

Risk management is usually performed through risk management frameworks, processes that provide procedural and prescriptive guidance on how to conduct risk management. The US NIST has provided an excellent general framework for AI risk management ([NIST, 2023](#)). ISO/IEC have published a guidance for risk management: ISO/IEC 23894:2023 ([ISO/IEC, 2023a](#)).

#### **Risk acceptability in the context of healthcare**

Every technology brings risks and while these can be minimized to a certain extent, risks can in most cases not be fully eliminated: there will always be *residual risks*. This triggers the question whether or not these are *acceptable*. Risk acceptability is a *relational* concept, i.e. it is very difficult to establish absolute thresholds for acceptability of (residual) risks. Instead, risks can be weighed against benefits (benefit-risk ratio) ([Pöyhönen, 2000](#); [Coglianese, 2020](#)) to establish whether residual risks are acceptable. The first requirement of the “Essential principles of safety and performance of medical devices” recommended by the IMDRF ([2018](#); see also [GHFT, 2005](#)), captures this:

*“Medical devices and IVD medical devices should achieve the performance intended by their manufacturer and should be designed and manufactured in such a way that, during intended conditions of use, they are suitable for their intended purpose. They should be safe and perform as intended, should have **risks that are acceptable** when **weighed against the benefits** to the patient, and should not compromise the clinical condition or the **safety of patients**, or the **safety and health of users** or, where applicable, **other persons**.”*

Thus, for AI-enabled products in healthcare, risks should be reduced to the extent possible, without however “adversely affecting the benefit risk ratio” (see Annex I of the EU's medical devices Regulation: “General safety and performance requirements”, Chapter I, point 2; [EU, 2017](#); [EU MDCG, 2020](#)).

To estimate the benefit-risk ratio, there is thus need for an evidence-based assessment of benefits (→ **BENEFICENCE**) to determine whether the residual and unavoidable risks are acceptable in view of the potential or real-world benefits.

Information on the benefit-risk profile of a medical device is collected during → **clinical evaluation**. Acceptability of risks should also consider the current state of the art (SOTA). See for instance Annex I, Chapter I, point 2 of the EU's medical devices Regulation, [EU, 2017](#). For a definition of SOTA, see → **added value**. For comments on SOTA versus benefit-risk ratio see Fraser et al., [\(2023\)](#) and Ebers [\(2024b\)](#). For an overview of a classification of benefit - risk profile *methods*, see [Najafzadeh et al., 2015](#).

## Explanatory note

### **General risk management for medical devices**

AI systems used in healthcare will typically be regulated as medical devices. It is therefore recommended to consider relevant guidance on essential principles (see [GHTF, 2005](#); [WHO, 2003](#)). For a comprehensive overview of many aspects relating to risk management for medical devices (including benefit-risk assessment), see also [Elahi \(2022\)](#). The ISO standard 14971 ([ISO, 2019](#)) “*Medical devices — Application of risk management to medical devices*” provides terminology, principles and an outline of process requirements. For a detailed proposal of a risk management process, see [Elahi, 2022](#): chapter 13—risk management process.

### **Risk management approaches beyond organisations**

ISO has provided a very general definition of risk management, which is targeted to organisations: “*Risk management refers to coordinated activities to direct and control an organisation with regard to risk*” (Source: [ISO, 2018b](#): ISO 31000:2018).

AI solutions in health are also produced and made available by non-organisational actors, including small research teams of researchers at universities, clinical settings or other research environments. The research community developing AI tools would benefit from heightened sensitivity to issues concerning risk management, including the adoption of a proactive and forward-looking approach to potential risks early during the research and development process (see list of AI-specific safety challenges under → **AI safety**).

### **Legal requirements for risk management systems**

The EU AI Act requires that, for high-risk AI systems, a risk management system needs to be set up and should be continuously and iteratively run throughout the life cycle (→ **life cycle of AI in health**). In broad terms, this system should allow for the identification and elimination or reduction of risks as far as possible and to allow to judge whether or not residual risks associated with specific harms or the overall AI system are acceptable ([EU, 2024a – Article 9](#)).

For AI products in health other legal requirements may be applicable, e.g. EU's medical devices Regulation ([EU, 2017a](#)), Annex I (General safety and performance requirements, Chapter I (General requirements), point 3: “*Manufacturers shall establish, implement, document and maintain a risk management system. Risk management shall be understood as a continuous iterative process throughout the entire lifecycle of a device, requiring regular systematic updating...*”.

## Term relationship:

Related terms:

- [AI safety](#)
- [NON-MALEFICENCE](#)
- [AI governance](#)
- [AI management](#)
- [AI evidence pathway for health](#)

# Risk

**Cluster:** B.2 Ethics, AI ethics, governance, management

## Concept description

### **Risk definition**

The term **risk** is usually understood as the product of (see for instance [NIST, 2023](#)):

- a) a specific harm or adverse event** associated with a product, system or service, and
- b) the likelihood (probability) of this harm/adverse event to occur.**

The **magnitude** of the harm/adverse event needs to be considered for risk estimations. Risk probability estimations need to be performed on the basis of the device, product, system or service being used as intended.

However, there may be also requirements to consider reasonably → foreseeable misuse (see for instance the provisions of the EU's MDR, Annex I, Chapter I, point 3; [EU, 2017a](#)). It is useful in this context to consider also variations in → use context and/or → use environment which can be reasonably expected and may have a bearing on risks. Inevitably, **risk estimations will be associated with uncertainty**. For that reason, robust risk estimations will, to the extent possible, draw on **reliable evidence** generated through the → AI evidence pathway for health.

### **Who or what is at risk**

When considering who or what is at risk, it is useful to structure risks broadly into the following groups, while keeping in mind that this categorization is not necessarily stringent, i.e. there may be overlaps:

- **Risks for individual persons** (→ AI safety; → NON-MALEFICENCE): safety risks for patients, users and other persons (e.g. healthcare professionals, users, family members using AI systems in home care etc.). Safety of healthcare professionals during execution of their duties relates to the concept of occupational safety.
  - **Risks for organisations, groups, communities and society as a whole.** This includes for instance risks of wrong diagnosis for specific patient groups that were not sufficiently considered when developing an AI system, either due to issues of → FAIRNESS (see → Avoiding discrimination and discriminatory bias) or on the basis of → universal versus targeted design. Other risks may stem from issues related to → health equality & health equity, the presence of → unavoidable trade-offs or the insufficient consideration of specific (vulnerable) groups (e.g. consider 'off-label' use of diagnostic devices due to the lack of dedicated devices for specific rare diseases (see → taking vulnerable groups into account "by design"). Risks include also research communities that adopt a solution that is affected by bias, thus biasing research results), risks for the society (e.g. due to the way AI solutions are adopted in healthcare and public health). There are also risks for healthcare professionals, e.g. job loss (→ job impacts) or → deskilling).
  - **Risks for property and assets of the value chain**  
Risks may also originate from value chain elements (e.g. enabling technologies), including distributed responsibilities across the value chain ([Cerna, 2018](#)). Cybersecurity poses a major risk to value chain assets / values (see [Reina & Griesinger, 2024b](#) for an ontology of cybersecurity, value chain enablers including IT infrastructure, enabling technologies as well as health impacts relating to cyber security incidents).
- Relevant risks include
- financial and reputational risks for a company (e.g. due to cyber attacks that affect assets of the value chain including model poisoning with subsequent deterioration of AI system performance and/or malfunction).

- risk for property including intellectual property (e.g. model theft). For data theft, see section below on risks relating to ethical principles.
- **Risks for the environment.** risks stemming from problems concerning resource sustainability (e.g. heightened energy and resource use), pollution and damage of natural resources (see → **SUSTAINABILITY**).

### **AI-associated risks can be largely related to ethical principles**

AI-related risks can, to a large extent, be related to ethical principles: see →**NON-MALEFICENCE**.

#### **Risk reduction and acceptable risks**

Technologies, especially those that directly affect persons and their health (e.g. medical devices), have intrinsically safety risks. These may be acceptable as long as expected specific benefits are deemed to outweigh these risks. Thus, in particular in healthcare, risk management cannot be completely disconnected from evidence on benefits (→ **BENEFICENCE**). See for instance the provision of the EU's medical devices regulation concerning risk reduction: Annex I, General Requirements, Chapter I, point 2: “...to reduce risks as far as possible means reduction of risks as far as possible without adversely affecting the benefit-risk ratio”; EU, 2017). Thus, given benefits, residual risks may be justifiable and deemed acceptable. See → **AI risk management** for more details.

## AI safety

**Cluster:** B.2 Ethics, AI ethics, governance, management

### Concept description

#### **1. AI safety**

Safety refers to the *state* of being protected from or unlikely to cause danger, risk, or injury. In health, safety relates to persons. **In the present context of AI in health, we understand AI safety as a state where harms and (in particular unacceptable) risks associated with AI are adequately controlled so as not to compromise the safety of patients, users and other persons.** From a → risk perspective, safety refers to a *state of minimized risks* for persons.

Relevant activities to minimise risks and achieve AI safety (→ **AI risk management**) may be part of general safety and risk management processes. While safety concerns persons, risks have a much broader scope, including also risks to property, the environment etc.

AI safety is closely related to → **AI risk management** and → **NON-MALEFICENCE**, the ethical obligation to avoid harm to people, groups, communities and society.

A comprehensive AI safety report of general AI risks has been prepared by a group of experts advising the UK government ([UK government, 2025](#)).

#### **2. Safety in health, clinical safety, health safety**

For medical devices safety has been defined as “Acceptability of risks as weighed against benefits, when using the medical device according to the manufacturer’s labelling” ([IMDRF, 2019 – Clinical Evaluation](#)). Thus, safety requires considerations of benefit-risk ratio, requiring an assessment of both → risks and of benefits (see → **clinical benefits for patient**; → **BENEFICENCE**).

Health technology (including AI systems in healthcare) should be as safe as possible, irrespective whether used in healthcare or for other health applications (e.g. research, health system management, public health surveillance). Safety in the context of health concerns:

- The **safety of patients** which are subjected, e.g. for diagnosis or treatment, to health products (including AI systems) as well as healthcare procedures (e.g. a given surgical procedure). Generally, safety means that health products do not cause harm or injuries if
  - a. correctly used (→ intended use) according to the → instructions for use and
  - b. used in the workflow or → use context (e.g. → clinical pathway) in which the product has been integrated to address a specific clinical problem.

However, safety needs also draw on considerations of reasonably → foreseeable misuse, which may include variations in → use context or → use environment.

- **Safety of users (e.g. healthcare professionals)** and **other persons** [WHO, 2003; GHTF, 2005](#). This aspect relates to occupational (health and) safety ([Niehaus et al., 2022; Fisher et al., 2023](#)).

AI safety corresponds to the terms → clinical safety and, what concerns patients, → patient safety. AI safety focuses however on the safety (and the specific safety challenges) of AI-enabled health products used in a clinical context.

### **3. Safety is typically achieved through both, quality and risk management**

Safety depends on the minimisation and control of risks, i.e. the probability of harms or adverse events to occur. As outlined, with AI safety we specifically refer to safety aspects in relation to AI-based elements of a product and/or procedure (e.g. robotic surgery) as opposed to the safety of the *entire* product. Thus, AI safety requires activities for identifying, understanding, analysing, minimizing and communicating those risks.

These activities may be carried out through → AI risk management (e.g. the establishment of structured risk management systems). Both the EU's AI Act ([EU, 2024a; Article 9](#)) and the EU's MDR ([EU, 2017a; Annex I, Chapter I, point 3](#)) require the establishment of risk management systems. Article 9 paragraph 10 of the EU's AI Act specifies that these can be combined if relevant conditions are fulfilled.

There is emerging technical guidance on → AI risk management: NIST has proposed an excellent general AI risk management framework ([NIST, 2023](#)). → NON-MALEFICENCE can be understood as relating to those elements of a risk management approach that are concerned with risks for people, i.e. patients (where applicable), users and other persons.

Importantly, risk management is closely linked to establishing and maintaining a quality culture within an organisation, e.g. through established quality assurance processes. These emphasise and put into practice the primacy of quality in regard to all activities and process steps within an organisation. From an organisational point of view, → risk management builds on quality assurance, ensuring in addition identification and control of potential harms and risks.

### **4. Specific safety challenges of AI-enabled health products**

**AI challenges the traditional notions of health product safety. Even if not completely unique to AI, there are several characteristics associated with AI technologies that have safety implications. These include:**

- **Risk due to AI being data based:** While all technologies depend on information and knowledge generated in underpinning scientific and technological research fields, AI products are *directly* based on knowledge in form of structured data. This introduces significant risks of → bias, e.g. due to incorrect data labelling. → Bias (systematic errors) can constitute a severe health hazard (consider for instance a diagnostic AI tool that shows representational bias and does not work properly for a specific ethnicity of patients). Bias may however not only enter through data, but also be based on the propagation of traditional biases in medicine in AI systems ([Straw, 2020](#)) or other incorrect assumptions (→ conceptual relevance), modelling decisions, evaluation metrics. Bias can be introduced at various points of the → algorithm-to-model transition.

- **Usability-related safety issues:** → Usability of an AI system may impact on the safety: poor usability may lead to errors and hazards.
- **Intelligibility:** Lack of → intelligibility of an AI systems' outputs may negatively impact on the decision-making in clinical contexts, leading to possible errors and safety issues. Lack of intelligibility has also implications for understanding failures and errors, which complicates the determination of root causes and may delay corrective actions (see explanatory note).
- **Distributed responsibilities** due to the complexity of the value chain. Consider for instance the common practice of acquiring data from data providers, models from model providers or platforms. This complicates risk management and distributes it on various actors. It also complicates → traceability, which is an important prerequisite for understanding failures and improving systems and for → failure transparency (external transparency).
- **Cybersecurity-related safety issues:** Due to the fact that AI solutions are depending on digital environments, there are significant cybersecurity-related risks, which may endanger patient health ([Reina & Griesinger, 2024a](#); [ENISA, 2023](#)), compromise the property and assets of companies or health settings, including data, models, AI systems as well as the functioning and integrity of (health) services.
- **The implementation and use of AI systems in real-world workflows may affect AI safety.** For instance, both automation bias and complacency (→ avoiding automation bias; avoiding → automation complacency) are potential safety risks ([Challen et al., 2019](#); [Arnold, 2020](#)). This includes also various drifts: → drift / shift in machine learning, in particular also → distributional drift / shift and → data drift / shift.

#### Explanatory note

**Safety**, effectiveness/efficiency and performance have to be systematically and rigorously addressed in the process of → **clinical evaluation** of medical devices ([IMDRF, 2019a](#)), a process which covers the pre- and post-market space.

Safety needs to be established prior to deployment of an AI system used in healthcare (e.g. through a clinical investigation) and needs monitoring once the AI system has been deployed for real-world use. This concerns post-market surveillance, PMS by the manufacturer and market surveillance by competent authorities. These surveillance activities may lead the detection of safety concerns and subsequent corrective actions, addressing their root causes.

What concerns post-deployment activities, we refer here to the definitions provided by the EU' Regulation on medical devices:

- “**Post-market surveillance** means all activities carried out by manufacturers in cooperation with other economic operators to institute and keep up to date a systematic procedure to proactively collect and review experience gained from devices they place on the market, make available on the market or put into service for the purpose of identifying any need to immediately apply any necessary corrective or preventive actions.”
- “A post-market surveillance plan should be developed and be readily available before the public and wide-spread use of the AI system in order to monitor the **safety** and **clinical performance** of the AI system in a real-world setting and timely detect any issues that may arise after the deployment.”
- “**Market surveillance** “means the activities carried out and measures taken by competent authorities to check and ensure that devices comply with the requirements set out in the relevant Union harmonisation legislation and do not endanger health, **safety** or any other aspect of public interest protection.”
- “**Corrective action** means action taken to eliminate the cause of a potential or actual non-conformity or other undesirable situation.”

## Term relationship:

Related terms: Text

- BENEFICENCE
- NON-MALEFICENCE
- FAIRNESS (concerning bias)
- TRANSPARENCY (e.g. failure transparency, traceability)
- Post-deployment monitoring
- Post-market surveillance
- Market surveillance
- Risk

## Human-centric AI

### Cluster: B.2 Ethics, AI ethics, governance, management

#### Concept description

Human-centric AI refers to a concept outlined in the European Commission's high-level expert group ethics guidelines for trustworthy AI ([EU HLEG, 2019](#)).

Human-centric AI is understood as follows:

*"The human-centric approach to AI strives to ensure that human values are central to the way in which AI systems are developed, deployed, used and monitored, by ensuring respect for fundamental rights, including those set out in the Treaties of the European Union and Charter of Fundamental Rights of the European Union, all of which are united by reference to a common foundation rooted in respect for human dignity, in which the human being enjoys a unique and inalienable moral status. This also entails consideration of the natural environment and of other living beings that are part of the human ecosystem, as well as a sustainable approach enabling the flourishing of future generations to come."*

#### Explanatory note

When using AI in medicine and healthcare, the concept of 'human-centric AI' concerns various ethical principles, notably → DIGNITY, FREEDOM AND AUTONOMY (e.g. potential detrimental changes to the patient-physician relationship. [Mittelstadt, 2021; Arnold, 2021](#)), environmental sustainability (→ SUSTAINABILITY).

A comparative glossary of terms has been compiled by the European Commission ([2022c](#)).

## Term relationship:

Related terms:

- AI ethics
- AI principles and AI ethics guidelines
- AI impact assessment (AI-IA)

## Research involving human subjects: ethical principles & guidelines

**Cluster:** B.2 Ethics, AI ethics, governance, management

### Concept description

Medical or clinical research aims at producing, through an optimised scientific study on a limited number of human subjects, knowledge on human health and biology that can be generalised to larger populations. Insights from clinical research enable better understanding of disease pathways and the development of novel diagnostic tools and treatments. Importantly, medical research ensures the safety and effectiveness of a given health product prior to its adoption, uptake and wide-spread use.

For health products (e.g. medicinal products, medical devices), such research is conducted through carefully planned studies that require ethical assessment and authorisation in order to protect the health of volunteers as well as scientific integrity. Ethical assessments of clinical research are based on ethical guidelines and codes. Important codes include: The Nuremberg Code (1947), the Belmont Report (1979), the Declaration of Helsinki (2004; last updated in 2024) (for an overview: [Kopjar, 2021](#); [WMA, 2024](#)). In addition, the Oviedo Convention ([Council of Europe, 2005](#)) contains an additional protocol concerning biomedical research, including interventions on human beings.

On the basis of the ethical principles and prescriptive moral statements of above mentioned and additional codes (e.g. [CIOMs \(2002\)](#) and US common rule ([US department of health and human services, 1991](#)), Emanuel et al. have suggested seven principles for ethical clinical research: social and clinical value, scientific validity, fair subject selection, favourable risk-benefit ratio, independent review, informed consent and respect for potential and enrolled subjects ([Emanuel et al., 2000](#); [US NIH, 2024](#)). These provide a clear framework for communicating research ethics involving human subjects.

Relevant study types in the context of this ontology include → **clinical investigations** on independent patient groups which may be part of → **model evaluation** and post-processing, aimed at evaluating *inter alia* → **bias**, → **interpretability** and assessing → **model performance** and → **model calibration**. This may serve to further improve models and AI systems prior to their broader use so as to avoid bias-based safety issues or insufficient insight into input-output relations.

### Explanatory note

#### Declaration of Helsinki

The declaration of Helsinki sets out ethical principles for medical research involving human participants. It was developed by the World Medical Association (WMA). Its final draft was accepted at a meeting in Helsinki in 1964 (latest version: [WMA, 2024](#)). Since then it has undergone eight revisions and two clarifications (for comments: [Bibbins-Domingo et al, 2024](#); [Reis, 2024](#) as well as [Shaw, 2024](#), commenting in particular on the **relevance of the Declaration in the context of the growing role of AI in health**; key changes of the 2024 include enhanced protection for vulnerable populations, improved transparency in clinical trials, and stronger commitments to fairness and equity in research).

It is the most widely accepted code of research ethics, making it a cornerstone for human research ethics. Fundamental principles of the Helsinki declaration are the right of individuals to make informed decisions regarding participation in clinical research (informed consent), the protection of vulnerable groups and individuals, the need for careful consideration of predictable risks and burdens over potential or foreseeable benefits to individuals willing to participate in clinical research, the precedence of participant's welfare over the interests of science and society as well as the precedence of ethical considerations over legislative frameworks.

#### The Belmont report

The report was created in 1978 by the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, the first public national body to shape bioethics policy in the United States. The Belmont report is a statement that summarises the basic ethical principles and guidelines for research with human subjects.

Respect for persons, beneficence and justice are the three core principles identified in the Belmont report while informed consent, assessment of risks and benefits, and selection of human subjects in research area, are stated as the three primary areas of its application.

**Oviedo Convention (Council of Europe)**

See → **DIGNITY, FREEDOM AND AUTONOMY**.

**Term relationship:**

Related terms:

- [AI ethics](#)

## B.3 AI actors and communities

### AI actors

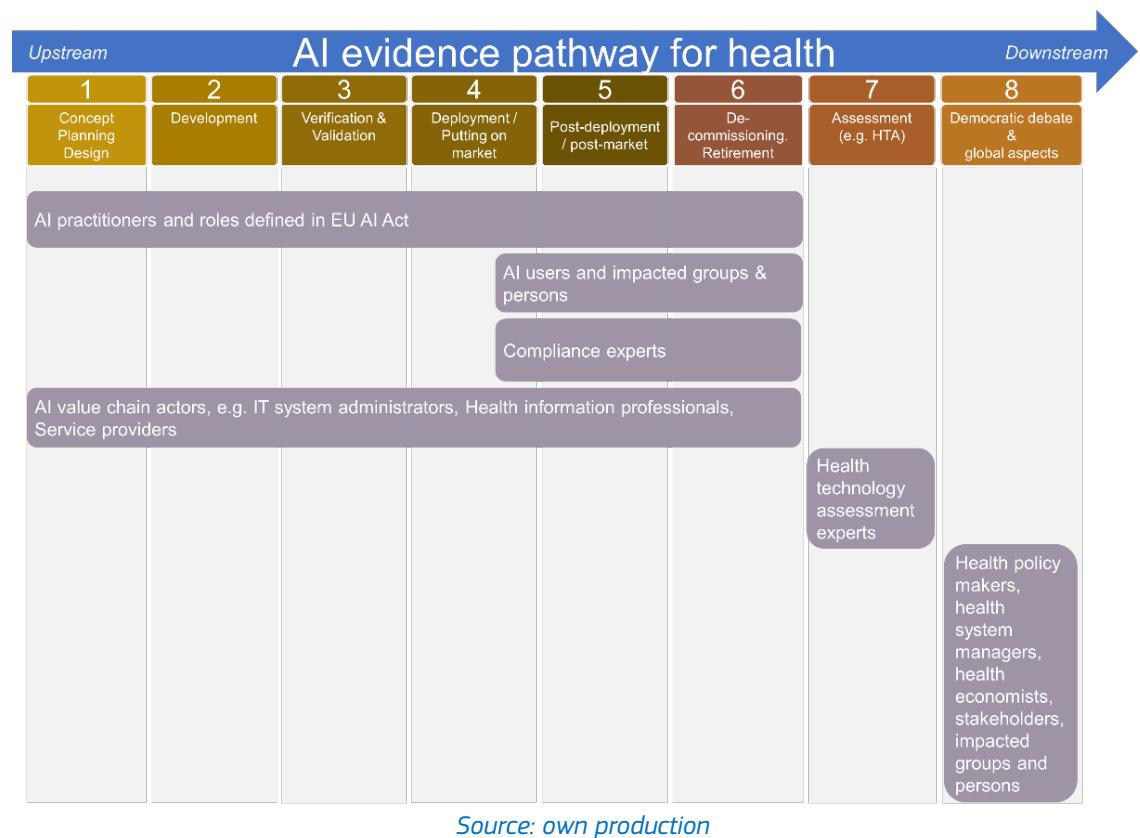
**Cluster:** B.3 AI actors and communities

#### Concept description

With AI actors we mean all people, organisations and communities of practice that have some form of agency, i.e. involved in activities concerning the development, deployment, proper functioning, monitoring and/or use as well as evaluation, appraisal and assessment of AI systems in healthcare and medicine as well as its implementation and integration into workflows, healthcare systems or public health tools. AI agents entail actors or stakeholders exerting influence via dialogue and feedback. Thus, AI agents in this broad sense encompasses:

- → AI practitioners  
(according to [EU HLEG, 2019](#))
- → Actors as defined in the EU's AI Act ([EU, 2024a](#))  
(providers, deployers, authorised representatives, importers, distributors and operators)
- → Users of AI in the health domain  
(e.g. researcher, healthcare professionals, patients)
- → AI value chain actors  
(e.g. health information professionals, IT system administrators, providers of services or models etc., hospitals engaged in telemedicine)
- → Compliance experts  
(e.g. experts at regulatory authorities and legislative bodies, experts at third party bodies (e.g. notified bodies), auditors)
- → Health technology assessment (HTA) experts
- → Health policy makers, health economists, health system managers

**Figure 20.** Schematic depiction of actor communities and their simplified relation in regard to the life cycle and AI evidence pathway in health



### Explanatory note

N.B. It should be noted that there is partial overlap between the terms → **AI practitioners** and → **Actors as defined in the EU's AI Act**. The latter covers (via “providers”) also organisations or people that develop AI. However, the definitions of the AI Act are from a perspective of legal responsibilities and therefore do not explicitly cover other relevant experts or professions (e.g. data scientists, AI system designers etc.).

We therefore use also the term → **AI practitioners** as suggested by the EU's high-level expert group.

### Term relationship:

Related terms:

- See concept description.

## AI practitioners

### Cluster: B.3 AI actors and communities

#### Concept description

The term AI practitioners means: “*By AI practitioners we denote all individuals or organisations that develop (including research, design or provide data for) deploy (including implement) or use AI systems, excluding those that use AI systems in the capacity of end-user or consumer.*” ([EU high-level expert group, 2019](#)).

Thus, in **healthcare**, AI practitioners include healthcare professionals but not patients using, for instance, AI-enabled wearables at home. 'Development' includes research into AI models, model creation, pre-training of algorithms, model provision (e.g. via platforms such as Github), data collection, data preparation or providing otherwise access to training data without data transfer (e.g. → **federated learning & split learning**).

AI practitioners in health and healthcare may be part of a complex non-linear → **value chain of AI**. Complexity here means that good practices of part of the practitioners do not necessarily result in improved outcomes across the value chain. AI practitioners constitute a multi-disciplinary community that requires collaboration to ensure trustworthy AI (→ **AI evidence pathway for health**).

In terms of relevant professions, AI practitioners in the health domain may include

- **Data scientists** (collecting, synthesising, wrangling data; conducting relevant impact assessments related to data use)
- **Health information professionals** (managing health data, including electronic health records)
- **Programmers**
- **AI system designers (experts)** identifying opportunities for AI systems, conducting user research in view of formulating the problem to be solved, collaborating with clinical/healthcare users, defining the use case (including → **use context** and → **use environment**), objectives and business model)
- **Clinical experts** in industry
- **Compliance experts** (e.g. working for regulatory agencies, notified bodies or auditing organisations, including clinical experts, IT experts, e.g. assessing cybersecurity of medical devices)
- **Experts on post-market surveillance**
- **Healthcare professionals using AI systems in clinical practice, typically in healthcare settings.** These also play an important role in ensuring performance in the post-deployment phase and alerting to potential drifts (or shifts) that may deteriorate model performance (→ **bias, heuristics, drift / shift**; → **monitoring drifts / shifts**).

In the areas of **health research, public health, and health system management**, fewer practitioners will be engaged (e.g. typically no involvement of compliance experts or healthcare professionals).

However, also for systems used in these domains, appropriate assessment of → **biases**, → **interpretability** and **explainability** and → **usability** (see also → **usability validation**) should be conducted. Equally **post-deployment monitoring** (e.g. through community discourse, peer review) is critical.

Finally, depending on the situation at hand, **experts in human rights impact assessments** may be required (e.g.→ **fundamental rights and algorithm impact assessment (FRAIA)**)

#### Explanatory note

N.A.

#### Term relationship:

Related terms:

- **AI actors**

## Actors as defined in the EU's AI Act

### Cluster: B.3 AI actors and communities

#### Concept description

In regard to the putting on the market of AI systems, relevant definitions can be found in legislation. We refer to the definitions of the EU's AI Act (Article 33, definitions 3 to 8; [EU 2024a](#)).

- 'provider' means a natural or legal person, public authority, agency or other body that **develops an AI system** or a general-purpose AI model or that has an AI system or a general-purpose AI model developed and **places it on the market or puts the AI system into service** under its own name or trademark, whether for payment or free of charge;
- 'deployer' means a natural or legal person, public authority, agency or other body **using** an AI system under its authority except where the AI system is used in the course of a personal non-professional activity;
- 'authorised representative' means a natural or legal person located or established in the Union who has **received and accepted a written mandate from a provider of an AI system** or a general-purpose AI model to, respectively, perform and carry out on its behalf the obligations and procedures established by this Regulation;
- 'importer' means a natural or legal person located or established in the Union **that places on the market an AI system** that bears the name or trademark of a natural or legal person established in a third country;
- 'distributor' means a natural or legal person in the supply chain, other than the provider or the importer, that **makes an AI system available** on the Union market;
- 'operator' means a provider, product manufacturer, deployer, authorised representative, importer or distributor.

#### Explanatory note

The AI Act outlines roles with specific responsibilities under the legislation. Other important actors (e.g. data scientists) are therefore not mentioned. The concept of → **AI practitioners** introduced by the EU Commission's high-level expert group in 2019 refers to a broader spectrum of agents.

#### Term relationship:

Related terms:

- **AI actors**

## Users of AI in the health domain

### Cluster: B.3 AI actors and communities

#### Concept description

**Users** refers to people that use AI systems in the four cardinal areas of AI use in health ([WHO, 2021](#)) for a specific application, need, → intended use and, where applicable, in a specific → use context or → use environment):

- health research
- health system administration
- public health surveillance and

- healthcare

Consequently, user in this context include:

- researchers, health system administrators (e.g. procurement, planning), researchers and officials active in public health surveillance and
- healthcare professionals including technicians (e.g. radiology), care staff and care givers of AI systems used in healthcare. This includes lay persons using → **AI-enabled medical device software**.

See → **user research** for definitions of user and lay person in the context of medical devices.

Importantly, users may be impacted when using an AI system. Impacts may be beneficial or adverse. The concept of user is understood to exclude active engagement in AI system development (see [EU, HLEG, 2019](#); see also other → **AI actors**).

- However, users, as stakeholders, may provide information and feedback on usability (→ **usability validation**) and user experience of an AI system, thus influence incremental changes.
- Further, users, through → **user research** activities may be indirectly involved in the specifications of AI systems (e.g. to be used in healthcare settings or for home care).
- Finally, while the role of user excludes AI system development, people or organisations may of course have several roles, e.g. clinicians may act as both, AI system developers and users in healthcare (→ **trust issues related to situations of conflict of interest**).

#### **Explanatory note**

In the context of safety and performance considerations of medical device technology, the term 'user' plays an important role.

Typically, there is a distinction of three groups whose safety needs to be ensured: 1) patients, 2) users (e.g. healthcare professionals, care staff) and 3) other persons ([see IMDRF, 2018; GHTF, 2005; WHO, 2003](#)).

#### **To summarise:**

- As outlined in the concept description above, the concept of **user of AI in the health domain has a broad scope**: it refers to users of AI tools in health research, public health, health system administration as well as healthcare.
- In contrast, the term 'user' in the context of safety of medical device technologies (see also → **AI safety**) **has a narrower scope**: it refers to **users of AI in healthcare only**, e.g. of → **AI-enabled medical device software**. These include healthcare professionals, care staff and care givers (e.g. family members), including lay persons (see → **user research** for definitions of 'user' and 'lay person' in this context).

A more detailed classification and subcategorization of users of medical device technology has been proposed by [Shah & Robinson \(2008\)](#).

#### **Term relationship:**

Related terms:

- **AI actors**
- **AI practitioners**

## AI value chain actors

### Cluster: B.3 AI actors and communities

#### Concept description

With value chain actors we refer to people involved in the various constituting elements of the → value chain of AI (see [Reina & Griesinger, 2024b](#))

Value chain actors include:

- **IT system administrators** that manage and maintain the **enabling IT infrastructure** and **enabling technologies** required for the development and deployment of AI systems. They supervise the installation, configuration, and maintenance of hardware, software, and network systems and ensure that the AI systems have the necessary resources and support for their development and operation. After deployment, they may be involved in monitoring system performance, troubleshooting and resolving technical issues that may arise. They may also be involved in implementing cybersecurity measures to protect assets of the → value chain of AI, e.g. data, → models, → AI systems. Additionally, IT system administrators may interact with developers to configure environments tailored to specific project needs.
- **Actors in the data domain** include a wide range of specialists, e.g. **data scientists** involved in collecting, processing and preparing data, **data providers**, **health information professionals** (HIPs).

The term health information professionals (HIPs) refers to persons that are involved in the design, development and implementation of healthcare information systems, including electronic health records (EHRs).

Healthcare information systems and EHRs may be maintained by a variety of (typically public) organisations, including hospitals, disease registries and databases.

HIPs are involved in the secure and ethically proper processing, storage and handling of sensitive medical information, including data pseudonymisation and aggregation (e.g. in collaboration with health researchers) and provision of health data to AI system developers under relevant applicable legislations and ethics codes.

- **Actors in the model and AI system domain** involve specialists that work on model development, AI system development and deployment as well as post-deployment activities.
- **Cybersecurity specialists:** cybersecurity as a horizontal ‘value-preserving enabler’ of the entire value chain of AI is critical for reducing vulnerabilities of specific assets and protecting value chain assets and enablers against attack and thus deterioration or destruction. This includes pertinent dangers such as breaches of personal data, data theft, model poisoning, model theft, interruption of services, and outage of services. In healthcare settings this may lead to patient health impacts ([Reina & Griesinger, 2024a, b](#)).

#### Explanatory note

##### **IT service administrators:**

In small companies/organisations ITAs cover a broad range of activities and responsibilities whereas larger companies/organisations may have separate positions with specific skills and tasks such as (see also [NIST, 2024b](#) for some of below terms):

- **Network administrator:** maintenance and management of the network infrastructure
- **Systems administrator:** maintenance and management of the IT infrastructure (e.g. servers, operating systems, hardware, user management)
- **Database administrator (DBA):** maintenance and management of databases including their performance, integrity, and security
- **Security administrator:** protection of IT systems and → data from unauthorized access, e.g. in the context of cybersecurity attacks that exploit vulnerabilities ([Reina & Griesinger, 2024a,b](#))

- **Cloud administrator:** maintenance and management of cloud-based infrastructure
- **Virtualization administrator:** management of → enabling technologies: platforms used to implement virtualized environments, such as virtual machines and containers.
- **Web administrators:** supervision of servers and web services

#### **Health information professionals:**

There have been various efforts towards certification of HIPs, in order to ensure accountability in line also with relevant ethical principles and legal provisions ([Gadd et al., 2016](#); [Kluge et al., 2018](#)).

#### **Term relationship:**

Related terms:

- AI actors

## Compliance experts

#### **Cluster:** B.3 AI actors and communities

#### **Concept description**

Experts directly involved in ensuring and supervising compliance of AI technology with legal and regulatory requirements and, in the EU, for the assessment of conformity with essential legal requirements ('conformity assessment').

Compliance experts include:

- Legislative experts, interpreting applicable law
- Experts at national authorities, supervising legislative implementation and the translation of legal requirements into regulatory pathways.
- Regulatory experts collaborating on international platforms (e.g. IMDRF) in view of harmonising global approaches with relevant for compliance.
- Experts at third party bodies such as notified bodies conducting conformity assessments.
- Third-party auditors.

#### **Term relationship:**

Related terms:

- AI actors

## Health technology assessment (HTA) experts

#### **Cluster:** B.3 AI actors and communities

#### **Concept description**

Experts that are involved in → health technology assessment (HTA). Given the multidisciplinary nature of HTA, HTA experts encompass various scientific, medical and economic disciplines and include also ethical experts.

#### **Explanatory note**

HTA experts should play a role in elaborating information requirements of clinical investigations of AI systems used in healthcare (see [Farah et al., 2023](#)).

**Term relationship:**

Related terms:

- AI actors
- Health technology assessment

Health policy makers, health economists, health system managers

**Cluster: B.3 AI actors and communities****Concept description**

- Experts involved in designing health policies and in analysing the effectiveness of healthcare.
- Experts managing, designing, assessing and improving health systems on a national, regional or supra-regional level.

**Explanatory note**

These experts play a key role in defining the way AI is implemented in health systems and in defining policies for governing AI in health systems on various levels.

**Term relationship:**

Related terms:

- AI actors

## B.4 Agency, autonomy and automation

### Agency

**Cluster:** B.4 Agency, autonomy and automation

#### Concept description

*"All other things must; man is the being that wills."*

Friedrich Schiller (in 'On the Aesthetic Education of Man in a Series of Letters').

Agency refers to the capacity of a person to act. Agency is closely linked to autonomy (or 'freedom from causation'): actors that are self-governing through application of their *free will* (e.g. the capacity to use reason to choose from a range of actions; see [explanatory note](#)) dispose of autonomy (for an introduction to autonomy, see [Ferrero, 2022](#)).

Agency and autonomy are central concepts in the philosophy of the European enlightenment (e.g. Immanuel Kant's concept of freedom as moral autonomy). Agency and autonomy both benefit from an actor that is able to make intelligent choices concerning his/her actions.

With the advent of AI, both concepts, agency and autonomy, are also applied to machines (e.g. AI systems, robots). This carries the danger of considerable confusion.

We therefore distinguish → **human agency** from → **machine agency**. We restrict here → **autonomy** to self-governing human or social agents (i.e. individuals or organisations) and use the term → **automation** to denote the extent to which machines or AI systems can operate without immediate human agency (e.g. supervision, oversight; see also → [ensuring human agency and oversight](#)).

#### Explanatory note

##### **Automation versus agency:**

One of the challenging aspects of AI systems and, indeed, one of their most promising ones is their capacity to execute complex ("intelligent") tasks with a minimum level of supervision (i.e. in an *automated fashion*; → **automation**; → **machine agency**), without fatigue and with accuracy and efficiency; this property is particularly interesting in healthcare: healthcare professionals use a considerable amount of their time for repetitive (e.g. administrative) work, for tasks that are straining and may be better tackled by a (supervised) AI system. The precise conditions and requirements of → **human agency** once an AI system has been deployed should be clearly described (see → [user competency and training requirements](#)).

##### **The concept of free will**

The concept of free will has occupied philosophers from antiquity to today. Kant formulated the question of free will and hence freedom from causation: "*whether pure reason of itself alone suffices to determine the will or whether it can be a determining ground of the will only as empirically conditioned.*" (Kant, Critique of practical reason). According to Kant, free will is an absolute freedom. Friedrich Schiller offered a nuanced modification "*By acting rationally at all man displays freedom of the first order; by acting rationally within the limits of matter, and materially under the laws of reason, he displays freedom of the second order*" (Schiller, On the Aesthetic Education of Man in a Series of Letters) (see [Noller, 2020](#)).

In the last century, neuroscience entered into the debate. Libet's findings on unconscious neural antecedents of motor behaviour ([Libet, 1983](#)) questioned the concept of free will (see however [Braun et al., 2021](#) for a review of Libet-style experiments). Newer findings show a middle ground between the two extreme positions, i.e. determinism and the postulate of absolute free will (e.g. [Lavazza, 2016](#)). "*Volition can be considered "free" as long as it is open to post-drive, second order processing ... characterized by an experiential sense of continuity. This may be seen as an alternative approach ... aligned with the neural antecedent prerogative.*" ([Dias, 2016](#)).

## Term relationship:

Related terms:

- Human agency
- Machine agency
- Automation

## Human agency

### Cluster: B.4 Agency, autonomy and automation

#### Concept description

We understand human agency in the present context as the capacity of a human actor to influence an AI system at various stages of the → AI evidence pathway for health, from concept, planning and design stages, deployment, integration into workflows, use to decommissioning ([EU HLEG, 2019](#)). Human agency is not dependent on whether an AI system is automated in regard to a specific task or not (→ automation).

Human agency is intended to ensure that AI systems are developed and used in a manner to allow appropriate control or oversight by human actors, in respect for fundamental human rights, universal values and ethical principles rooted in human dignity (→ human oversight; → ensuring human agency and oversight). Human oversight is central for responsible AI.

In the context of AI in healthcare, human agency, → human oversight and → human primacy of healthcare professionals will support trust in the patient physician relationship (→ upholding a trustful patient-physician relationship). Human agency should also help avoiding that the practice and art of medicine relies too much on machine interpretable outcomes and terms ([Mittelstadt, 2021](#)) with a de-humanising effect on healthcare as well as potential consequences on safety ([Ebers, 2024a](#)).

#### Explanatory note

Human and machine agency are seen by many researchers as increasingly entangled in complex "human-machine networks" ([Eide et al., 2016](#), [Tsvetkova et al., 2017](#)) or complex "assemblages" of machine and human agents ([Ananny & Crawford, 2018](#)). See also → affective computing.

While human agency (→ human oversight) should ideally ensure → human-centric AI and AI that serves humans, respects universal values (such as dignity and → human autonomy; → DIGNITY, FREEDOM AND AUTONOMY), involvement of humans in the design, development and post-deployment control does not guarantee that these values are indeed respected.

This requires that human actors adhere to ethical principles during all phases of the life cycle (AI design, use, post-deployment), which in turn is only then a realistic proposition if high-level ethical principles are broken down and translated into actionable aspects tailored to the use domain of a given AI system. It also may require regular auditing (→ auditability and auditing) not only in view of performance and safety, but also in regard to ethical aspects, including detection of discrimination or bias (see: [Mittelstadt, 2021](#)).

## Term relationship:

Related terms:

- Human primacy
- Human oversight
- Affective computing

## Human oversight

**Cluster:** B.4 Agency, autonomy and automation

### Concept description

Human oversight (i.e. control over the functioning and output generation of AI systems) is particularly relevant at the post-deployment stage, i.e. once an AI system has become operational within a given → use context and → use environment. Effective human oversight requires → corrigibility of AI systems.

There are challenges to human oversight: humans are not always reliable in fulfilling oversight tasks, e.g. due to lack of sufficient or adequate competence, or harmful incentivisation ([Laux, 2023](#)). Thus, human oversight requires

- that users have the knowledge and tools to comprehend AI systems and have sufficient knowledge about key design features
- that AI systems are intelligible (→ intelligibility), that their way of operating can be comprehended and that outcomes can be explained and, within limits, predicted based on the understanding of the system's workings. Intelligibility requires clear and understandable communication and documentation (→ intelligibility). Making AI systems intelligible has impacts on → traceability, → failure transparency and the capacity to challenge outcomes → contestability and challenge.
- That there are means to effectively interact with and control them, e.g. through appropriate user interfaces

Article 14 of the EU's AI Act ([EU, 2024a](#)) outlines requirements to ensure human oversight of high-risk AI systems.

Established oversight concepts include ([EU HLEG, 2019](#)):

- human-in-the-loop (HITL),
- human-on-the-loop (HOTL), or
- human-in-command (HIC).

### Explanatory note

N.A.

### Term relationship:

Related terms:

- Human agency
- Human primacy
- Corrigibility

## Affective computing

**Cluster:** B.4 Agency, autonomy and automation

### Concept description

Affective computing relates to the capacity of AI systems to detect human affections, feelings and sentiments ("sentiment analysis") and to process such data in a manner to generate outputs that appear like affective human behaviour ([IEEE, 2017](#)).

While affective computing may facilitate human-machine interfacing (HMI), there are also substantial risks for individual users to place excessive and misguided trust in an AI system (→ avoiding automation bias; → avoiding automation complacency).

#### Explanatory note

Affective computing may have considerable potential for use in medicine and healthcare to support the detection and/or treatment of mental or cognitive diseases / impairments and to support patients with specific risks (e.g. suicide).

Affective computing is used for instance in → **conversational agents**, in wearable (medical) devices ([Schmidt et al., 2019](#)), for mood and cognitive disorders ([Smith et al., 2021](#); [Sarris, 2022](#); [Sajno et al., 2023](#)).

There are also risks however, including over-bonding, aggravation of mental health issues, dehumanisation of patient care (see → **upholding a trustful patient-physician relationship**).

IEEE' document on ethically aligned design contains a section on affective computing ([IEEE, 2017](#)).

#### Term relationship:

Synonyms:

- Artificial emotion intelligence
- Emotion AI

Related terms:

- Cognitive computing

## Autonomy

### Cluster: B.4 Agency, autonomy and automation

#### Concept description

Autonomy in Western ethics and political philosophy means freedom of choice: the capacity of human actors to make choices when deciding on acts and the capacity to base these choices on personal reasons, personal values, aspirations and desires, which implies a potential tension with ethics and its central aim of leading a meaningful life by acting justly and not only for the personal good, but in respect also of other person's autonomy, values, choices. This has been formulated in various cultures as the 'golden rule' (Confucius, Thales, Mahabarata, Christianity, Islam), i.e. "do unto others as you would have them do unto you" ([Höffe, 1986](#)).

Autonomy is thus closely related to → **agency** and dignity. For example, in Kant's philosophy (in particular '*Groundwork of the metaphysics of morals*' and '*Critique of practical reason*') autonomy, personhood and dignity are closely intertwined (see → **DIGNITY, FREEDOM AND AUTONOMY**), with personal autonomy being at the root of *moral responsibility*. Autonomy can be seen a precondition of human freedoms and relates thus to concepts of fundamental human rights and freedoms.

We suggest that autonomy should not be used to denote machine → **automation**, or → **machine agency**, i.e. the property of machines and robots to fulfil specific tasks without human supervision or → **human oversight** (see also [European Commission, 2018](#): European Group on Ethics in Science and New Technologies, Statement on artificial intelligence, robotics and 'autonomous' systems).

#### Explanatory note

##### **Autonomy and resulting accountability**

The concept of autonomy can be extended to organisations which can, as an assemblage of individuals, make specific choices and, by doing so, can be held morally accountable for their choices and actions, in

particular in regard to potential tensions and conflicts with legitimate interests of groups or the society as a whole.

However, it is obviously meaningless to hold a machine or AI system morally accountable: it is ultimately the human actors that designed and deployed a machine or AI system that have moral accountability (see also [Roff, 2019](#); [Babushkina, 2023](#)).

#### ***Autonomy as misnomer for automation***

The term 'autonomy' is also used to denote grades of machine → [automation](#), which can include levels of automation without need for direct human interference for a given task during a given period. Nevertheless, the term 'autonomy' has crept into technical and everyday parlance to denote → [machine agency](#) and → [automation](#). Examples are:

- The use of the term "autonomous driving" to denote computer automation of driving personal vehicles, also referred to as "self-driving" or "driverless cars" (preferred terms); see information by [Union of Concerned Scientists](#). Online: <https://www.ucsusa.org/resources/self-driving-cars-101> or
- 2) The term "autonomous" robotic surgery (see for example: [Saeidi H et al. 2022](#)); the preferred term would be 'automated' or 'fully automated' robotic surgery.

For a comment on *autonomy* versus *automation*, see [Chiodo S \(2022\)](#).

#### **Term relationship:**

Related terms:

- [Human agency](#)
- [Human primacy](#)
- [Patient primacy](#)

## Human primacy

#### **Cluster: B.4 Agency, autonomy and automation**

#### **Concept description**

Under human primacy we understand the concept that AI systems should be and in fact are (so far) always ultimately controlled by human actors (→ [human agency](#); → [ensuring human agency and oversight](#)).

This holds also for 'fully automated' AI systems, since also such systems are designed, deployed, monitored and decommissioned by human actors.

In addition, systems that are not fully automated at their operational stage, will still need some level of human supervision (see also → [augmentation](#)).

#### **Explanatory note**

N.B. The term "human primacy" has not the same meaning as 'human-centric AI'. The latter is a term introduced by the European Commission's High-level expert group document ([EU HLEG, 2019](#)) and essentially refers to adherence to ethical principles and in particular human → [DIGNITY, FREEDOM AND AUTONOMY](#).

#### **Term relationship:**

Related terms:

- [Autonomy](#)
- [Human agency](#)

- Corrigibility
- Human-centric AI
- Patient primacy

## Corrigibility

**Cluster:** B.4 Agency, autonomy and automation

### Concept description

The concept that the output or behaviour of AI systems can be corrected, i.e. is *corrigible* by human intervention or human input during operation of the AI system in the post-deployment space ([Soares et al., 2015](#)). We consider corrigibility a translational concept of → NON-MALEFICENCE (see also → AI safety).

Corrigibility means to constrain → machine agency and → automation to a certain extent by enabling human input as an “active veto” ([Vergheze et al., 2018](#)) at any time of operation. Corrigibility is a necessary condition for effective → human agency and → human oversight; designing AI systems with corrigibility safeguards is essential for maintaining → human primacy.

In that sense, corrigibility relates to the concepts of “augmented intelligence” in medicine (→ [augmentation](#)). While corrigibility is a necessary precondition for the safe operation of AI (e.g. by correcting misdiagnoses or erroneous clinical recommendations), corrigibility is no safeguard against automation complacency (→ [avoiding automation complacency](#)).

### Explanatory note

Corrigible AI systems can be defined as systems that “tolerate, cooperate or assist many forms of outside correction.” ([Firt, 2024](#)). A key motivation for ensuring corrigibility by design is the growing use of complex → [foundation models](#) and → [generative AI](#), trained on a broad spectrum of (multi-modal) data (→ [data modality](#)) and resulting in a broad-spectrum of potential functionalities and applications. This complexity may however create unforeseen outcomes and it is hence critical to adapt and correct such complex systems in view of their → [intended use](#) within a given → [use context](#) or → [use environment](#) before they are deployed and maintain the possibility to correct such systems once deployed.

### Term relationship:

Related terms:

- Autonomy
- Human agency
- Human oversight
- Human-centric AI
- Patient primacy
- Augmentation

## Patient primacy

**Cluster:** B.4 Agency, autonomy and automation

### Concept description

The primacy of the patient over societal, scientific and economic interests, intended to protect the dignity, identity and integrity of patients.

This concept is spelled out in the Council of Europe's "Oviedo Convention" which promotes the protection of human rights in biomedicine at a transnational level ([Council of Europe, 1997](#)).

The Oviedo convention builds on principles and rights of other human rights conventions and treaties, notably the

- *United Nations' International covenant on civil and political rights of 1966* ([United Nations, 1966](#))
- *Council of Europe's European Convention on human rights of 1950* ([Council of Europe, 1950](#)), in particular in view of rights to life, physical integrity and privacy (→ PRIVACY PROTECTION), prohibition of discrimination (see → FAIRNESS and → DIGNITY, FREEDOM AND AUTONOMY).

### Explanatory note

- a) The European Convention of human rights of 1950 should not be confused with the later issued Charter of fundamental rights of the European Union ([EU, 2000](#)). However, the EU's Charter, mapping out fundamental legal rights is a bedrock of EU legislation, including on digitisation (e.g. relevant provisions of the General Data Protection Regulation or the Data Act).
- b) For a detailed discussion on the potential impact of AI on patient's rights and in particular the patient-physician relationship in the context of the Oviedo convention and the European convention on human rights, see: Mittelstadt, B ([2021](#))

### Term relationship:

Related terms:

- Human-centric AI
- PRIVACY PROTECTION
- FAIRNESS
- DIGNITY, FREEDOM AND AUTONOMY
- Respecting patient primacy

## Machine agency

**Cluster:** B.4 Agency, autonomy and automation

### Concept description

The capacity of an artificial system (e.g. an AI system, a robotic system) to act and/or interact with its physical environment, thereby causing changes to that environment.

### Explanatory note

Note that machine agency simply denotes the capacity of a machine to act; it does not necessarily imply that the machine is fully automated ("autonomous") for a specific task or range of tasks.

The extent of independent autonomous machine agency ("machine autonomy") is captured under the term "automation". Note that a concept similar to 'machine agency' which we posit here, has been previously named 'object agency' (Jia et al., 2012) in the context of discussions on the internet of things (IoT).

Machine agency and human agency are seen by many researchers as increasingly entangled in complex "human machine networks" (Eide et al., 2016, Engen et al., 2016) or complex "assemblages" of machine and human agents (Ananny & Crawford, 2018)

#### Term relationship:

Synonyms:

- AI actors

## Automation

#### Cluster: B.4 Agency, autonomy and automation

#### Concept description

Automation refers to the capacity and extent to which machines can operate independently from → **human agency**, at least what concerns a given defined task or range of tasks.

IBM (2024) defines automation as follows: "*Automation is the application of technology, programs, robotics or processes to achieve outcomes with minimal human input.*"

#### Explanatory note

While there are possibilities for automation also in medicine and healthcare (e.g. in health system administration, management, procurement or robotic surgery), fully automated systems need to be implemented with prudence so as not to affect human dignity.

Even full automation for a specific task does not imply an absence of human agency and oversight at stages preceding and after deployment of an AI system, e.g. human agency at design, production and deployment stages and human oversight as the capacity of interfering with "task-autonomous" AI systems.

The review by Mennella et al. (2024) provides an overview of possible levels of automation of AI in healthcare.

#### Term relationship:

N.A.

## Augmentation

#### Cluster: B.4 Agency, autonomy and automation

#### Concept description

Augmentation refers to capacity of AI systems to augment human agency, focusing on the assistive role of AI rather than its potential to replace human actors. Augmentation thus means to enhance the quality, efficiency, performance and relevance of human actions and → **human primacy**.

While augmentation is not the same as → **automation**, efficient augmentation may involve a certain degree of → **automation**. In medicine "augmentation" refers to the use of AI under strict conditions of →

human oversight and → human agency (see for instance the Statement of American medical association on augmented intelligence in medicine).

#### Explanatory note

Examples for augmentation are decision-support systems in clinical care, support of image segmentation tasks or radiology-based diagnostics.

For augmentation in healthcare see Crigger et al., (2022) and American Medical Association (2024, updated 2025).

#### Term relationship:

Related term:

- Automation

## B.5 Bias, heuristics, drift & shift

### Bias

**Cluster:** B.5 Bias, heuristics, drift & shift

#### Concept description

Bias in the context of AI can refer to:

- 1) systematic error (this entry) or
- 2) to → neural network bias which, in simple terms, refers to a constant added to the product of features and weights in → artificial neural networks (e.g. convoluted neural networks (CNN), widely used for visual AI tasks in medical image analysis).

#### ***Bias = systematic error***

Bias (or “systematic error”) can be defined as the “*systematic distortion of results or findings from the true state of affairs, or any of several varieties of processes leading to systematic distortion. In everyday usage, “bias” often implies the presence of emotional and/or political prejudices that influence conclusions and decisions.*” ([Last, 2007: dictionary of public health](#)). Bias is distinct from “random error”, which primarily affects overall → precision. While undesirable, random error is not unfair, although this may a priori seem so during *ex post* evaluations. Bias however may be discriminatory or unfair in cases where the error systematically affects persons or groups characterised by specific → features or → attributes.

#### ***Two fundamental aspects of bias***

In the context of AI use in health and medicine, we distinguish two fundamental aspects of bias:

- ***Algorithms may be biased:*** Since AI has → machine agency and can be employed to support decision-making, bias in AI is a major concern. Biases may creep into AI systems throughout the → algorithm-to-model transition, leading to → algorithmic bias. Examples of such bias are systematic errors in classifying sub-groups of patients ([Obermeyer et al., 2019](#)) or estimate levels of risk of disease. Groups that are underrepresented in medical research (e.g. due to ethnicity or disease profile) are particularly susceptible to bias ([Vokinger et al., 2021](#)). There are strategies for mitigating bias (e.g. [Nazer et al., 2023](#)).
- ***People may be biased.*** Biased → human agency clearly impacts how people (e.g. physicians) act under specific circumstances ([Fitzgerald & Hurst, 2017; Bagnis et al, 2021](#)). This concerns various “cognitive biases”, especially in the context of heuristic decision-making (→ **heuristics**). Thus, the clinical implementation of AI systems may introduce biases of various sorts ([DeCamp & Lindvall, 2020](#) [Mittermaier et al, 2023](#)), including automation bias and complacency (→ **DIGNITY, FREEDOM AND AUTONOMY**) ([Arnold, 2021; Challen et al., 2019](#)). Since there is agreement that medical consequential decision making should not be fully automated (e.g. [AMA, 2024](#) or [United Nations, 2021](#)<sup>43</sup>), human bias (e.g. through → heuristic decision-making) should be considered. It should be noted that AI supported consequential decision making in healthcare may lead to a reduction of heuristics-based and biased decisions (→ **BENEFICENCE** → **clinical gains**).

<sup>43</sup> See paragraph 123(e): “Member States should pay particular attention in regulating prediction, detection and treatment solutions for health care in AI applications by... (e) ensuring the human care and final decision of diagnosis and treatment are taken always by humans while acknowledging that AI systems can also assist in their work”

Both aspects can work together to perpetuate bias with negative impacts on health equity. For an overview of most common sources of bias in healthcare AI and proposed mitigation measures, see: [European Parliament, 2022](#).

### **Bias and fairness**

Bias in artificial intelligence is typically understood as lack of → FAIRNESS and thus also to health equity ([Rajkomar et al., 2018](#)). Data sets or a model may not be fully balanced (and thus be biased or skewed) in regard to various → features which may lead to consistent outcomes that however *consistently* disfavour a sample.

Bias of AI in medicine is of course not only an issue of fairness, but also of clinical safety since (inequitable) bias may harm patients, e.g. by favouring consistently false negative results of a diagnostic tool when applied to a specific patient group of a specific ethnicity ([Obermeyer et al., 2019](#)).

Bias detection should be an active process during the entire transition from algorithm to model (→ **algorithm-to-model transition**). Since data are a major source of bias, procedures, standards and other tools may help to address bias in a consistent manner ([Griesinger et al., 2022; NIST, 2022](#)).

### **'Intrinsic' biases**

Perhaps most difficult to detect are intrinsic biases, e.g. biases that are encapsulated in historical data sets or in historical or traditional assumptions or views that build the conceptual foundation of a model (→ **conceptual relevance**). Further complication of bias detection stems from pre-trained models, which may contain intrinsic biases that are 'hidden' in the connectivity-determining → **parameters** (e.g. weights) of an artificial neural network and may only be detected through sophisticated explainability techniques or during → post-deployment monitoring or → post-market surveillance (→ **risks related to insufficient post-deployment monitoring / post-market surveillance**). Such biases may be detected through audits that should address performance, safety as well as biases and, in particular, potential discrimination of specific individuals or groups (→ **accountability**; - > **auditability and auditing**).

### **'Latent' biases**

So-called 'latent biases' ([DeCamp & Lindvall, 2020](#)) are biases that evolve *after* clinical application. Sources of latent biases might involve **i) adaptive learning**: the continuously learning model integrates over time uncorrected biases that exist in a given novel environment or from another context or even within the health system as such (→ **continuous and adaptive learning**); **ii) biases related to human agency**, e.g. automation bias and automation complacency (→ **DIGNITY, FREEDOM AND AUTONOMY**); **iii) biases in regard to the choice of 'outcome of interest'** that does not really reflect interests of patients or communities. We consider this as a bias of → **conceptual relevance**.

### **Residual bias**

It should be noted that, in practice, it is nearly impossible to eliminate bias completely. For instance, the fact that it is not possible to satisfy simultaneously various notions of fairness or justice in a model (→ **Intrinsic incompatibilities or 'trade-offs'**) could be considered unavoidable residual bias. However, it is critical that such known bias is transparently described and that all effort is undertaken to reduce bias, particularly where it may represent a risk (→ **AI safety**; → **NON-MALEFICENCE**). Risk reduction should however not negatively affect the benefit-risk ratio.

### **Equitable bias**

While inequitable bias can harm patients (→ **FAIRNESS**; → **AI safety**; → **NON-MALEFICENCE**) ([Ranard et al., 2024](#)), but so-called "equitable bias" has been proposed as a desirable feature under specific circumstances and for specific → **use contexts**, e.g. to tweak an AI system in a way so as to ensure that underrepresented population groups or patients with rare diseases are sufficiently considered ([Tejani et al., 2024a](#)).

### **Ethical dimensions of bias:**

To conclude, bias has a series of ethical dimensions and associated more specific issues (**Figure 21**):

- Biased AI (→ **algorithmic bias**) can be a safety issue (→ **NON-MALEFICENCE**).
- Healthcare professionals integrating AI into workflows can introduce various forms of bias when using AI. Automation bias and automation complacency of healthcare professionals and

health systems as a whole can undermine the dignity of patients and the patient-physician relationship (→ DIGNITY, FREEDOM AND AUTONOMY).

- Biased AI may be unfair, discriminatory, non-inclusive and lead to disparate outcomes for people with sensitive attributes (→ FAIRNESS).
- Biased AI may become an issue of → contestability and challenge, lead to → remedy and redress requests or even litigation under → liability law (→ RESPONSIBILITY), e.g. in situations where bias led to harm of patients and/or users (→ NON-MALEFICENCE; → AI safety).
- Finally, bias may concern specific vulnerable groups / patient groups that are not sufficiently considered when new AI solutions are being conceived and designed (→ SOLIDARITY).

**Figure 21.** Schematic depiction of the ethical dimensions of bias and associated issues. Orange font: bias of people.



Source: own production

#### Explanatory note

Suresh and Guttag (2021) have proposed a useful categorisation of biases in AI. We use this framework to briefly outline potential biases in the context of health and medicine:

- Historical biases: All bias that is based on historical data, assessments or scientific concepts and views that are outdated, scientifically incorrect (e.g. an association between a biomarker x and a disease y, that turned out a spurious correlation, are flawed or were tarnished by judgements that are sexist, racist or “blind” in regard to specific aspects of sub-populations or disease groups).
- Representation bias: A bias that is based on a mismatch between the data used to train a model and the realities of the population or environment in which the model will operate so that the training data do not properly represent that population / environment, thus compromising the performance of the AI model if no precautions are taken. For instance, when developing an AI system for predicting the risk of ischaemic heart disease, only insufficient data from specific groups (e.g. women, Asians etc.) are available. Thus, skewed representation of the real-world situation in the training data, but also lack of specific data points can be considered representation bias. Representation bias, if properly understood and documented, can be captured in descriptions of → applicability and limitations.
- Measurement bias: measurement bias occurs for instance when specific → proxies are being used that turn out to be oversimplifications of a more complex situation. Consider for instance the use of the proxy “family history” for predicting susceptibility to cancer. Giving such a proxy too much weight in a predictive approach may obfuscate other risk factors, such as life-style, weight, habits, exposure to chemicals etc.
- Aggregation bias: this type of bias can occur when different subsets of data are integrated or aggregated into one set of data, using uniform approaches that are based on assumptions of consistency of the datasets that are however flawed.

- Learning bias: this bias type stems from modelling choices that amplify disparities of data, e.g. with underrepresented attributes.
- Evaluation bias: This bias occurs when the data used in the testing (or test-benchmark) phases do not represent the actual use population. Testing or benchmark data that are characterised by misrepresentations will propagate such misrepresentations during phases of model optimisation, leading ultimately to the development and deployment of models that are “tailored” to work well on the subset of testing / benchmark data, but are not sufficiently representative for the real-world situation.
- Deployment bias: can originate when a model is used differently than foreseen during its development. An example is the use of a model for prediction as specific disease that is used in the → use environment also for predicting another (e.g. rare) disease. While the model may be helpful in this context, the model design intended for another use scenario may bias outcomes of the “off-label” use. Note that deployment bias is also referred to as → **distributional drift / shift**.

#### Term relationship:

Synonyms:

- Systematic error
- Disparate impact

Related terms:

- **Algorithmic bias**
- **Heuristics**
- **FAIRNESS**

## Heuristics

### Cluster: B.5 Bias, heuristics, drift & shift

#### Concept description

Heuristics refers to a way of making decisions that draws on experience, precedence and preference. It is also used as a concept to reflect a specific view on human rationality in regard to making decisions. Heuristic decision-making is an important topic in medicine and healthcare, where decisions often have to be taken under time-pressure ([Whelehan et al., 2020](#); [Islam et al, 2014](#); [Wegwarth et al., 2009](#)).

Ideally, human decision making would rely on a thorough analysis of all pieces of evidence or, at least all elements of information at hand, weighing these information sources according to specific criteria (e.g. of quality, completeness, adequacy, relevance) and then integrate all information in order to produce a decision in response to a specific problem. This hypothetical decision-making process is referred to as “unbounded rationality” or “optimisation” ([Marewski & Gigerenzer, 2012](#); [Aliouche, 2022](#)). However, such decision-making is under real-world conditions not possible or practicable.

Instead, according to a school of thought in cognitive science, people make “heuristic” decisions, based on experience, precedence or by giving specific information sources strong weight – without always having sufficient reasons for doing so. This does not necessarily imply that heuristically reached decisions are wrong. They may represent simply a mental “shortcut” that has been found, by experience, to work under given circumstances: consequently, heuristics has been defined “*a simple procedure that helps find adequate, though often imperfect, answers to difficult questions*”<sup>1</sup> ([Kahneman, 2011, cited in Hjeij & Vilks, 2023](#); for references on the ‘research programme on heuristics and biases, see [Miller & Gelman, 2018](#)). In cognitive science and some schools of philosophy, heuristics is understood as making decisions based on “cognitive biases”, e.g. simple rules, previous experiences in a similar situation

("availability heuristic"), precedents or specific information that is held to be of particular relevance for making appropriate decisions in a given situation ("anchoring heuristic", e.g. a specific diagnostic tool overriding consideration of other information in a given clinical decision making situation).

Marewski & Gigerenzer (2012) distinguish three views of human rationality in relation to decision making: *unbounded rationality: optimisation* and the heuristic approaches of a) *irrationality: cognitive illusions and biases*, c) *ecological rationality: fast and frugal heuristics*.

Finally, it is conceivable that patterns of heuristic decision making may also incorporate → **bias**. Thus, heuristics may also touch on → **FAIRNESS** and → **NON-MALEFICENCE**.

<sup>1)</sup> Already Pierre-Simon de Laplace wrote in his *Essais philosophique sur les probabilités* ([Laplace, 1825](#)): "One sees in this essay that the theory of probabilities is basically only common sense reduced to a calculus. It makes one estimate accurately what right-minded people feel by a sort of instinct, often without being able to give a reason for it".

## Explanatory note

### **Human heuristic decision-making**

Heuristic decision-making plays an important role in medical practice and in particular in clinical settings, given the various pressures faced by clinicians (time, efficiency, patient overflow, economic pressures etc.). Despite its real-world importance, it seems that heuristics receives surprisingly little attention in discussions of practices of healthcare and medicine and this includes novel perspectives on possible heuristic decision-making associated with the use of AI systems.

AI in healthcare may, as a benefit, help → **reducing heuristic decision-making** by enabling faster and more accurate data processing, integration and interpretation, thus enhancing the quality of evidence use for clinical decisions and potentially reducing → **bias** from consequential decision-making in clinical care. AI may also enable a higher degree of shared decision-making (SDM) ([Abbasgholizadeh Rahimi et al., 2022](#)), enabling less paternalistic models of patient care (→ **DIGNITY, FREEDOM & AUTONOMY**).

However, if AI solutions are taken up in an uncritical manner, they may become an element of heuristic-decision-making (through the "anchoring" heuristic), that is giving an AI system overly high weight for a specific decision without sufficiently weighing in other facts and sources of information ([Mosier & Skitka, 2006](#)). This concept is closely related to automation bias and complacency (→ **avoiding automation bias**; → **avoiding automation complacency**).

### **'Heuristics' in machine learning**

Optimisation procedures of → **artificial neural networks** have been called heuristic: while powerful, humans cannot immediately understand how they work ([Lipton, 2018](#)). In addition, neural networks can exhibit so-called → **shortcut learning** (see also → **artificial intelligence (AI)**). Shortcut learning happens when the algorithm learns specific recognisable patterns, that are however not key decision features. The resulting outputs may be correct on benchmark data, similar to the → **training data**, but may be incorrect under situations that deviate (even slightly) from the training situation. Thus, outputs appear accurate during training, validation and testing but may in fact not be robust and generalisable (→ **generalisability**).

## Term relationship:

Related terms:

- BENEFICENCE – clinical gains: reducing heuristic decision-making
- Avoiding automation bias
- Avoiding automation complacency
- Clinical practice protocol
- Clinical practice guideline (CPG)
- Clinical pathway
- Cognitive psychology
- Behavioural economics

## Shortcut learning

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

### Concept description

#### **Origin of the term: shortcut learning in neurosciences and cognitive psychology**

Shortcut learning in cognitive psychology and neurosciences refers to the acquisition of mental shortcuts, whether consciously (i.e. learning) or subconsciously, for a variety of cognitive tasks, including recognition, judgements, conclusions, explanations etc. Such shortcuts represent simpler rules that may be based on the correlative presence of specific properties that are associated with the primary property of interest. In the context of decision making, shortcut ‘rules’ are used for →heuristics-based decision making.

#### **Shortcut learning in deep learning-based machine learning**

In the context of → deep learning-based machine-learning, shortcut learning refers to a corresponding phenomenon, i.e. the training data-based extraction of features that are easily recognisable but not fully relevant for the problem to be addressed (e.g. recognition of tumours in radiological images) (Geirhos et al., 2020). A machine learning model may for instance associate a specific irrelevant anatomical structure that is correlative present in tumour-positive images with the outcome tumour-positive, i.e. produce accurate (→ accuracy) recommendations, that are nevertheless based on irrelevant features or features with insufficient predictivity and of no → conceptual relevance. The available literature on shortcut machine learning provides a variety of examples (Hermann et al., 2024) of trained models preferring various properties over the properties of interest, e.g. textures over shapes or background (colour) over the actual foreground objects of interest (see the example in Yam, 2025).

#### **Shortcut learning is rooted in the ‘economics’ of model optimisation**

It appears that models can under circumstances privilege *availability* of features (i.e. how easily they can be extracted from data) over *predictivity* (i.e. how reliably a feature indicates training set labels) (Hermann et al., 2024). While this is sometimes termed ‘laziness’ of or ‘cheating’ by models, it is important to recall that, mathematically, → machine learning approaches simply favour the simplest solutions in terms of their → objective function, i.e. by applying relevant optimisation constraints, model developers may favour the ‘quickest’ route to maximising performance metrics.

#### **Detecting shortcut learning**

Shortcut learning is not always evident on the basis of → validation data or → testing data or with other benchmark data that are similar to those used during → model development. However, shortcut learning for visual AI can be detected through the ‘**occlusion sensitivity**’ approach (Petsiuk et al., 2018; Valois et al., 2023), where, successively, various parts of an image are occluded to determine what drives a given output and detect potential irrelevant features associated with outputs (this approach can be considered an explainability approach). Shortcut learning may moreover become apparent once the model is faced with real-world data or data that differ from the property domain of its training data, including through audits (→ auditability and auditing). Lack of → generalisability (as evidenced by insufficient accuracy for data out of the property domain of the → development data) may be an indication of shortcut learning but could also point to more general problems (e.g. → overfitting).

#### **Shortcut learning: a severe limitation and risk of neural network / deep-learning based approaches**

Shortcut learning obviously poses a considerable risk for the safe and reliable use of → machine-learning models. If shortcut learning-based decisions go undetected prior to deployment of a model, the model may make incorrect decisions under more general real-world conditions, which, in healthcare, might lead to incorrect diagnostic or clinical decisions with consequences for → patient safety. Shortcut learning is thus an issue of → AI safety and → NON-MALEFICENCE (see → risk associated with insufficiently controlled short-cut learning). Adequate quality-check processes including external → model validation and → model evaluation in smaller contexts prior to routine deployment may help detecting shortcut learning issues.

### **The risk of shortcut learning is best controlled by striving for intelligibility of models**

The best approach for controlling shortcut learning is, quite clearly, to strive for → intelligibility of the model, i.e. an understanding why and how it makes decisions, allowing to understand whether it makes the right decisions for the right reasons.

#### **Term relationship:**

Related terms:

- Artificial neural networks
- AI technique
- Heuristics
- Intelligibility
- Interpretability and explainability

## Algorithmic bias

#### **Cluster: B.5 Bias, heuristics, drift & shift**

#### **Concept description**

The totality of biases intrinsic to a trained algorithm or model (for a short introduction, see: [National library of medicine, 2024](#); [Awan, 2024](#); [IBM, 2024](#)). Algorithmic bias may be introduced by insufficient → conceptual relevance (see also → valid clinical association / scientific validity) or by various decisions made along the → algorithm-to-model transition.

Algorithmic bias can be a consequence of various data-related → biases (e.g. historical, representational, measurement, aggregation, learning, evaluation, deployment; choice of → features, → attributes, → proxies etc.), but may also be caused by modelling decisions (→ machine learning model).

Algorithmic bias is a major concern when using AI systems in clinical settings ([Saint James Aquino, 2023](#); [Mittermaier, 2023](#)).

#### **Term relationship:**

Related terms:

- Bias

## Intrinsic incompatibilities or ‘trade-offs’

#### **Cluster: B.5 Bias, heuristics, drift & shift**

#### **Concept description**

Intrinsic incompatibilities refer to objectives that cannot be achieved simultaneously by a model or AI system for scientific and/or technical reasons. Such incompatibilities lead to ‘trade-offs’, i.e. situations where developers need to balance these objectives and, ultimately, make decisions which objective has priority given the intended application purpose of the model.

Examples for intrinsic incompatibilities:

- In drug development, there may be conflicts between various objectives such as maximising structural novelty versus maximising structural similarity to molecules with known bioactive properties ([Zhang et al., 2025](#)).

- AI-based test systems that generate predictive classification outcomes (e.g. dichotomous, two class predictions like ‘positive’, ‘negative’) can either be optimised for higher sensitivity or higher specificity, but cannot satisfy both at the same time. This constitutes an ‘intrinsic incompatibility’. For instance, when designing a medical diagnostic test, decisions must be made on whether the test should be optimised for detecting the *positive group* (i.e. people with having the disease being correctly classified as positive) or for the *negative group* (people not having the disease being correctly classified as negative). This implies trade-offs: optimising for → **sensitivity** (true positives) means trading off against misclassification of people that do not have the disease (false positives) ([Char et al., 2020](#)). Such optimisation or → **model calibration** decisions may also have ethical implications since, depending on calibration, a given test may be seen as disadvantaging a given group.

### Explanatory note

Kleinberg et al., (2016) have elaborated on this topic under the title "incompatibility theorem" in relation to → **FAIRNESS**: specific notions of justice or fairness in the context of probabilistic classifications are incompatible with each other, i.e. they cannot be satisfied simultaneously, resulting in inevitable trade-offs between different notions of fairness, irrespective of the specific use context and also irrespective of the method used for producing the probabilistic classification.

Kleinberg et al distinguish three notions in relation to a dichotomous or binary classifications:

- 1) Calibration within groups,
- 2) Balance for the negative class,
- 3) Balance for the positive class.

Only one of these calibration criteria can be satisfied, no statistical model can satisfy all three at the same time.

This is relevant for a variety of applications, including classification recommendations in the context of health, healthcare and diagnosis. The impossibility to satisfy all three criteria means that AI developers need to make a conscious and conscientious decision, ideally with end users and stakeholders, in regard to which ‘calibration’ criterion is deemed most important within a given use context to achieve outcomes that are as fair as possible and as useful as possible with the ultimate aim to ensure that potential sensitive → **attributes** (e.g. gender, sex, race, age etc.) do not interfere with the classification of the model / AI system. See also the explanatory note for → **algorithmic bias**.

### Term relationship:

Related terms:

- FAIRNESS
- Model development
- Model calibration
- Bias

## Drift / shift in machine learning

### **Cluster:** B.5 Bias, heuristics, drift & shift

### Concept description

“Drift” or “shift” in AI / machine learning refers to the undesired phenomenon that the performance of a model may deteriorate during the post-deployment stage due to an evolution of the environment in which the model is used (e.g. [Quinonero-Candela, 2008](#); [Mardziel 2021](#); [Kagerbauer et al., 2024](#)).

There are various scenarios and root causes that can lead to so-called model or algorithm drift: 1) → concept drift / shift; 2) → data drift / shift (also referred to as ‘covariate shift’) and 2) → distributional drift / shift (also called ‘deployment bias’).

#### Term relationship:

Synonyms

- Model drift
- Algorithm drift

Related terms:

- Concept drift / shift
- Data drift / shift
- Distributional drift / shift

## Concept drift / shift

#### Cluster: B.5 Bias, heuristics, drift & shift

#### Concept description

Concept drift refers to a change over time of the statistical properties of the target variable which a model is trying to predict. Patterns and ‘rules’ or associations learned by the model during the training phase are not any longer fit for purpose. As a consequence model performance (e.g. accuracy) may degrade. Such → concept drift / shift is primarily an issue in situations where the underlying data distributions may change abruptly ([Klaiber et al., 2023](#)).

A recent example of concept drift concerns the COVID-19 pandemic: there is evidence that specific AI systems for clinical care (e.g. for mortality and sepsis prediction) degraded in their performance ('performance drift') (e.g. [Rahmani et al., 2023](#); [Parikh et al., 2022](#)).

Continuous and adaptive learning is used as a strategy to tackle challenges of → concept drift / shift. However, this may involve risks: continuous learning may allow → bias (from new or changed environments) to creep into models (see → bias; “latent biases”).

#### Term relationship:

Synonym:

- Drift / shift in machine learning

## Data drift / shift

#### Cluster: B.5 Bias, heuristics, drift & shift

#### Concept description

Data drift occurs when the historical data (e.g. retrospective health data) which were used for training the model, are not any longer representative of the current situation or conditions. Thus, there is a drifting apart of → post-deployment input data from the → training data used during model development; these are no longer ‘representative’ of the real-world problem ([Quinonero-Candela et al., 2009](#); [Bialek, 2024](#); [Ali, 2023](#)).

This includes distribution of data and/or data quality (e.g. pixel resolution of medical images). To establish that potential data drift is not obfuscated by other drifts, it is important to ascertain, in cases of

suspected data drift, that neither the input-output function (which could happen in cases of continuously learning models) nor the way the model is being used (→ **distributional drift / shift**) have changed.

#### Explanatory note

Finlayson et al. (2021) have proposed a checklist for recognising and mitigating dataset shift in clinical settings, clustered among the topics i) technology, ii) changes in population and setting, iii) changes in behaviour.

#### Term relationship:

Related terms:

- Drift / shift in machine learning
- Distributional drift / shift

## Distributional drift / shift

**Cluster:** B.5 Bias, heuristics, drift & shift

#### Concept description

Distributional shift /drift (also called ‘deployment bias’\*; Suresh & Guttag, 2021) can originate when a model is used differently than foreseen during its development, e.g. changes in clinical practice that affect the way the data produced by the AI model are being used (Finlayson et al., 2021).

\*) see the explanatory note of the entry → bias.

#### Explanatory note

A specific example is the use of a model for prediction of a specific disease that, in addition, is used in the → **use environment** for predicting another (e.g. rare) disease. While the model may be helpful in this context, the model design intended for another use scenario may bias outcomes of the “off-label” use.

#### Term relationship:

Synonym:

- Deployment bias (see → bias)

## B.6 AI Evidence pathway, AI life cycle and AI value chain

### AI Evidence pathway for health

**Cluster:** B.6 Evidence pathway for health

#### Concept description

The AI evidence pathway for health is a conceptual framework focusing on **evidence generation** for trustworthy AI in health through **collaboration of communities along the life cycle and value chain**, covering the pathway **from conception to adoption**. It is intended to foster **safe and secure innovation** by ensuring **bridging interdisciplinary communities** ('community bridging') to identify upfront necessary evidence required at subsequent stages of the pathway and by generating relevant evidence for trustworthy AI (→ TRUST AND TRUSTWORTHINESS).

The pathway emphasizes the a) role of people for evidence generation and evidence use ("people centric"), b) the need to orchestrate several elements to ensure robust evidence and c) the collaboration of relevant agents & communities.

It is rooted in an ontology of AI concepts in health, composed of A) ethical principles and translational concepts to bridge towards granular technical, scientific and clinical requirements and B) related fundamental AI concepts in regard to health applications.

The five elements required for effective collaboration towards evidence for trustworthy AI are:

- 1) **Agents and communities** that make decisions on requirements, needs and design, make AI and deploy, use, monitor, decommission, evaluate and assess AI solutions in health (see → AI actors and communities).
- 2) **Life cycle stages** that focus on specific processes aimed at developing, producing, deploying, monitoring, evaluating and decommissioning AI systems (see → life cycle of AI in health). We propose<sup>44</sup> a life cycle comprising two additional stages as compared to the most frequent proposals in the literature. These two stages are important for safe AI innovation and successful deployment of AI technologies on markets. For healthcare uses, they influence AI conception and design, e.g. through decisions on how AI systems are used in clinical workflows or pathways:
  - **the systematic socioeconomic evaluation of AI-enabled health products** through 'health technology assessment (HTA) (see also → life cycle of AI in health) and
  - **broader debates, e.g. on the way AI is implemented in clinical practices, the future use of AI in healthcare** (see → DIGNITY, FREEDOM AND AUTONOMY - AI and the development of healthcare), the **equitable roll-out of AI in countries and regions** (see → FAIRNESS – health equality & health equity) and global aspects linked to AI technologies including the risk of creating new colonial and exploitative structures (see → SOLIDARITY – avoiding colonial structures; reducing the global north-south divide).

This responds to principles of international public bodies that do not directly relate to ethical principles, such as "democratic values" ([OECD, 2019a, amended 2024](#)), "public interest" ([WHO, 2021](#)) and "safe innovation" ([Council of Europe, 2024](#)).

<sup>44</sup> There is neither consensus on a general AI life cycle nor on one specific to health.

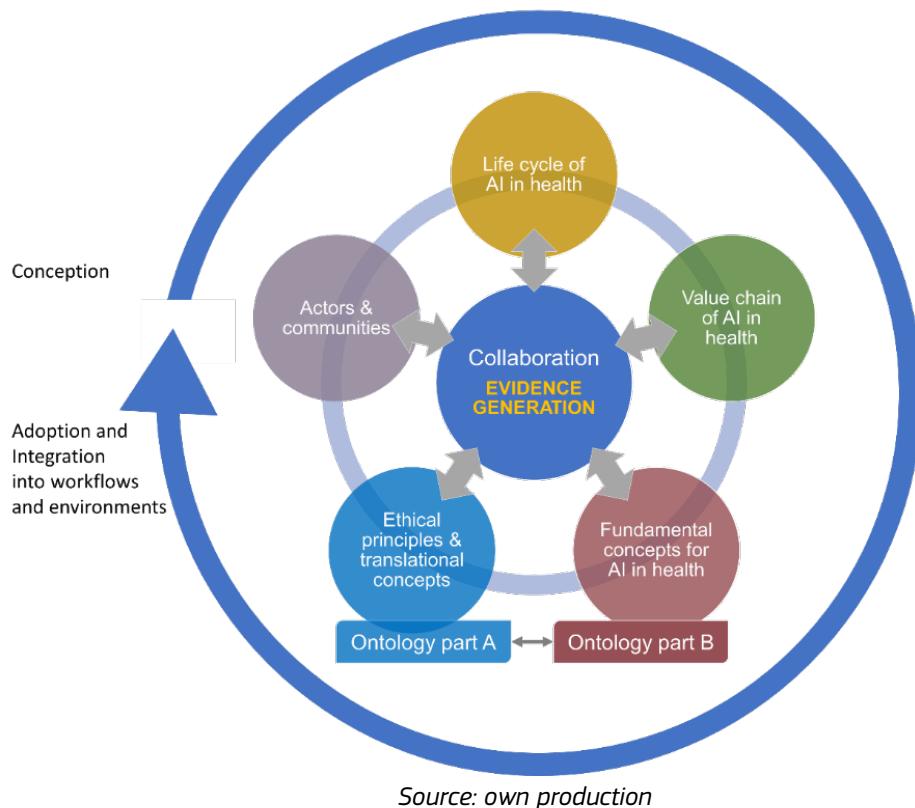
- 3) Value chain elements that are prerequisites for AI.** We distinguish “**enablers or prerequisites**” from “**values or assets**”, with cybersecurity as a fundamental ‘value-preserving’ enabler of all other value chain elements (see → value chain of AI in health).
- 4) Consensus ethical principles and lower-level translational concepts** for bridging AI ethics and life cycle processes (**Part A of this ontology**: Figure 22- blue rectangle);
- 5) Fundamental AI concepts for AI in health** to support common understanding and collaboration amongst agents and communities (**Part B of this ontology**: Figure 22 - red rectangle).

#### Explanatory note

**Figure 22.** The AI evidence pathway is a conceptual framework for interconnecting all relevant elements for creating evidence for trustworthy AI.

The five elements contributing to the AI evidence pathway are shown: **1) Actors and communities** that define requirements and needs, make decisions and specifications, deploy, use, monitor and supervise, decommission, evaluate and assess AI solutions in health; **2) Value chain elements** that are required for the development, production, deployment and use of an AI system. We distinguish ‘assets’ (e.g. data, models, AI system) and ‘enablers’ (e.g. Enabling IT infrastructure); **3) Life cycle stages** that focus on specific processes aimed at developing, producing, deploying, monitoring, evaluating and decommissioning an AI system; **4) 9 granular ethical principles** with lower-level translational concepts to be used along the life cycle. These are described in part A of this ontology (blue rectangle); **5) Description of fundamental concepts** relevant to AI in health with references to health-relevant publications to support common understanding and collaboration amongst actors and communities. These are described in part B of this ontology (red rectangle). Other tools in preparation are shown as light grey rectangles.

## AI evidence pathway for health



## Term relationship:

Related terms:

- Life cycle of AI in health
- Value chain of AI
- Actors and communities
- AI ethics
- Ethical principles
- Ethical evaluation of AI
- Bioethics
- Principlism

## Life cycle of AI in health

**Cluster:** B.6 Evidence pathway for health

### Concept description

#### **Life cycle: overview:**

The term life cycle denotes the passing of an item (e.g. a physical product, a process, a conceptual framework etc.) through subsequent stages from conception to broad availability. The term is related to the concept of value chain (→ value chain of AI), but not identical.

The life cycle imagery of biology has been adopted in various contexts, including **development of products and processes**. It captures the **process of transiting from the conception of a solution for a specific problem to the tangible solution**, typically involving a **marketable product or service**, as well as the **iterative and cyclical aspects of redesigning a product** based on the experiences made during its real-world use.

- Depending on need, a life cycle may contain various “stages” that relate to specific activities. A typical life cycle is **process-oriented** and **less evidence-oriented**. It outlines tasks and activities that need to be done at various stages of developing a product or process. However, **evidence requirements** that would satisfy the needs of various communities are not a primary concern of life cycle processes. We therefore suggest to integrate life cycle approaches in the broader framework of the → AI evidence pathway for health.

#### **Various proposals for an AI life cycle, but no consensus:**

For AI **various life cycle stages have been proposed**. While there are usually common elements, there is currently **no consensus concerning the stages an AI ‘life cycle’ would comprise**. It is also questionable whether such a definition would be useful if not targeted to the specifics of an application domain, e.g. health (see explanatory note).

#### **Proposal for a life cycle for AI in health (Figure 23):**

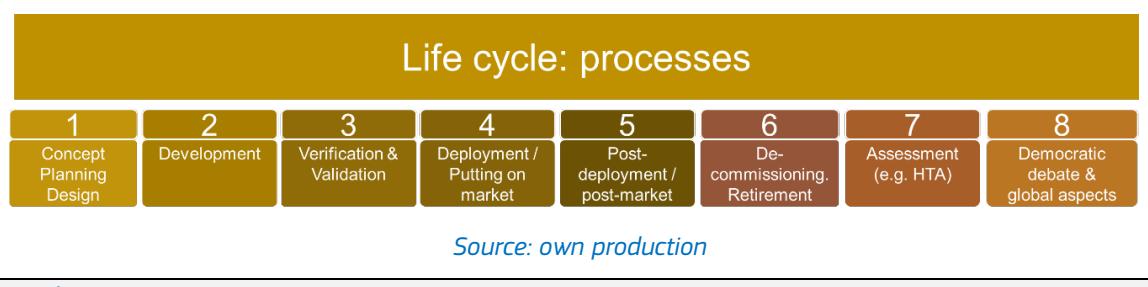
##### 1. **Concept, planning & design:**

- Formulation of the problem to be addressed by the AI solution, including the identification of current gaps regarding solutions and market opportunities (for products/services) or needs (e.g. for solutions within a scientific community)
- Gathering evidence on → conceptual relevance of the intended AI system. In case of systems used in healthcare: evidence on → valid clinical association / scientific validity
- Framing of → intended use, presumptive → applicability and limitations, possible → use context and → use environment considerations, including limitations and use restrictions.

- Design of the solution based on requirements of the intended users (→ user research)
2. **Development:** Development of the AI system. This involves the algorithm / model (→ algorithm-to-model transition; → model development), other relevant components, interoperability considerations regarding the → use environment and → use context.
3. **Verification and validation** of the solution including, iterative steps of improvement. This stage may include verification of the AI system design specifications, validation of the model used in the AI system (→ model validation), → AI system validation ('analytical / technical validation'), → usability validation. For healthcare AI solutions, these activities may be part of → clinical evaluation.  
We subsume under this stage also → model evaluation and post-processing activities prior to full → clinical validation. Model evaluation helps assessing a model on the basis of independent patient groups, checking for residual → bias that may compromise accuracy and safety, evaluating the degree of → intelligibility, → interpretability and explainability. The latter are important for informed patient consent (→ ensuring the means for free and informed consent) and for understanding and tracing errors (→ traceability) which is essential for monitoring and improving AI solutions.
4. **Deployment / Putting on the market:** This encompasses
- Transiting from development and research settings towards production at scale and establishment of quality-controlled production processes.
  - Deployment or putting on the market of the solution, including description of → intended use, → instructions for use, → applicability and limitations, required IT infrastructure needs, enabling technologies and other value chain enablers like cybersecurity ([Reina & Griesinger, 2024b](#)) as well as required → user competency and training requirements for proper use in line with the → intended use, intended → use context and → use environment.  
Description of residual → bias and possible → intrinsic incompatibilities or 'trade-offs'.
- N.B. For high-risk devices that are regulated under relevant legislations, e.g. EU's AI Act, sufficiently detailed documentation needs to be made available on key aspects of the AI system** (see EU AI Act, Article 11 on 'technical documentation'; [EU, 2024a](#)) and relevant other legislation for specific documentation needs (e.g. EU's medical devices regulations; [EU, 2027a](#)).
- Generally, consideration should be given to sufficient → disclosure of additional elements that may support trust.
5. **Post-deployment / post-market.** Depending on the use of the AI system this may involve
- For non-healthcare AI systems (e.g. health research): → monitoring usability and usability-related errors. Updating, de-bugging, corrective actions etc. Monitoring of performance and possible drifts / shifts (→ monitoring performance, effectiveness, efficiency and bias; → drifts / shifts).
  - For AI systems used in healthcare → post-market surveillance including → incidents / → adverse events and AI-specific aspects: Monitoring of → bias that may only become evident once the system is deployed and used under real-world conditions; monitoring for drifts / shifts, notably → distributional drift / shift, → data drift / shift; supervision of how residual → bias and → intrinsic incompatibilities or 'trade-offs' affect clinical performance and safety; monitoring of → real-world benefits and possible unforeseen risks; communication with user communities on performance, safety, usability and possible improvements of the AI system.
6. **Decommissioning or retiring** the AI system. Due to the specifics of AI solutions such as sensitive and confidential health data generated and stored within a given → use environment, → decommissioning / retirement needs careful attention.

7. **Assessment (e.g. HTA):** If the solution concerns an AI system used for healthcare, an important stage concerns an independent socioeconomic, clinical and ethical assessment, e.g. through a process of → health technology assessment (HTA).
8. **Democratic debate & global aspects:** This stage is not part of a product-specific life cycle, but developments concerning the use of AI in healthcare or aspects of its global societal impact may feed back into upstream life cycle activities of any given product.

**Figure 23.** Proposed life cycle stages for AI in health. Stages 1-7 concern specific AI solutions. Stage 8 concerns broader discussions which however may radiate back to life cycle activities of individual products. A life cycle should be seen in the larger context of → AI governance, → AI management including → AI risk management activities. The → AI evidence pathway builds on this life cycle outline, connecting it to the → Value chain of AI and enabling community collaboration for evidence generation of trustworthy AI.



#### Explanatory note

While life cycle definitions of AI tend to be highly similar, there is no consensus on the precise staging of an AI life cycle. For instance, the life cycle definitions of the UN, OECD or the US centers of excellence are similar but not fully aligned (e.g. disassembly / termination is only included by UN):

- Life cycle ranges “from research, design, and development to deployment and use, including maintenance, operation, trade, financing, monitoring and evaluation, validation, end-of-use, disassembly and termination of AI.” ([United Nations, 2024a](#)).
- “An AI system lifecycle typically involves several phases that include to: plan and design; collect and process data; build model(s) and/or adapt existing model(s) to specific tasks; test, evaluate, verify and validate; make available for use/deploy; operate and monitor; and retire/decommission. These phases often take place in an iterative manner and are not necessarily sequential. The decision to retire an AI system from operation may occur at any point during the operation and monitoring phase.” ([OECD, 2019a, amended in 2024; see also OECD 2024a](#)).
- “The AI lifecycle is the iterative process of moving from a business problem to an AI solution that solves that problem. Each of the steps in the life cycle is revisited many times throughout the design, development, and deployment phases.” ([US government – Centers of excellence AI, 2024](#)).

In addition to proposals put forward in documents of international organisations, various valuable concepts for an AI life cycle have been proposed in the scientific literature:

- De Silva & Alahakoon and have noted that there is not consensus on life cycle definitions for AI: various descriptions of AI life cycles and methodologies “either do not render comprehensive coverage from conception to production, or are limited in the level of detail of each individual phase” prompting the proposal of a detailed life cycle for AI, comprising 17 distinct stages ([De Silva & Alahakoon, 2021](#)).
- Ng and colleagues have proposed an AI life cycle for ethical AI in the context of health decisions, comprising five stages from data creation to model deployment ([Ng et al., 2022](#)).

- Similarly, Abràmoff and colleagues have proposed a so-called “*Total Product Lifecycle (TPLC) framework*” for machine learning / AI healthcare applications (‘ML-HCA’) ([Abràmoff et al., 2023b](#)), based on an earlier paper by Char et al. which focused on a bioethical analysis of the AI life cycle including identification of relevant stages from conception to deployment and implementation ([Char et al., 2020](#)).
- In above mentioned paper, Char and colleagues have proposed a framework composed of a life cycle (‘development & implementation pipeline’) and outlined also work streams such as evaluation and oversight, decisions and characteristics, that raise value-based issues as well as ethical considerations ([Char et al., 2020](#)).
- Similarly, NIST\_in its AI risk management framework ([NIST, 2023](#)), integrates a life cycle approach with a parallel stream of identifying actors and relevant activities, focusing however more on risk identification, risk framing and management and less on a broader ethical evaluation based on ethical principles.
- A common understanding of the term life cycle and its stages would support a consistent and coherent application of life cycle-based approaches for designing and assessing AI systems in medicine and healthcare. In any case, a life cycle approach for AI development should not lead to the deferring of evidence generation to the post-market phase as has been identified as a problem in the area of medical devices ([Harkin et al., 2024](#)).

#### **Term relationship:**

Related terms:

- AI evidence pathway for health
- Value chain of AI
- Health technology assessment (HTA)
- Decommissioning / Retirement

## Value chain of AI

#### **Cluster: B.6 Evidence pathway for health**

#### **Concept description**

The terms → **life cycle** and **value chain** are often used without clear definition and may even be used interchangeably ([UK department for environment, food and rural affairs, 2017](#)). The term ‘value chain’ analysis was introduced by Porter in 1998 to delineate the principle activity of a sector of business ([Porter, 1998](#)), dissecting the underlying activities a business performs in designing, producing, marketing, and distributing its product or service. Life cycle in contrast is a sequence of subsequent phases through which an item (e.g. product) passes ([Falkowitz, 2025](#)).

We define value chain as the *physical prerequisites, process prerequisites and knowledge prerequisites* within one or several organisations and their sequential or parallel connectedness (“chain”) for creating items, products or processes with new and added value in comparison to the state of the art. This definition draws on a text by the [UN \(2024\)\\*](#).

In the current context, the values created are models, AI systems and its associated services for using AI in health (e.g. for telemedicine using → **AI systems** or → **AI-enabled medical device software**). We further distinguish ‘enablers’ or prerequisites from ‘assets’ or values within the value chain.

Our proposed value chain of AI in health has the following dimensions:

#### **Enablers / prerequisites:**

- 1) Enabling IT infrastructure, e.g. computers, storage (cloud local), networks etc.
- 2) Enabling technologies, e.g. machine learning libraries, platforms (containerisation, orchestration) etc. This may include also services providing specific products, including pre-trained models.
- 3) **Cybersecurity** competence as a horizontal value preserving enabler, protecting all other value chain elements from deterioration / destruction / theft through cyber attacks
- 4) **Data**, e.g. health data from electronic health records, imaging data for development of diagnostic models, textual data for AI systems supporting clinical knowledge integration and decision making

#### **Assets / values:**

- 5) **Models**\*\*, i.e. machine learning models that are developed *de novo* or have been developed through → **transfer learning** based on pre-trained (typically large) models
- 6) **AI systems**, incorporating AI models and, depending on case, other components, such as a user interface, actuators, sensors etc.
- 7) **Services** associated with using the AI system in health

Notably, data, models, AI systems and services vary depending on specific AI application, while enabling IT infrastructure, enabling technologies and cybersecurity are not dependent on the specific case.

**Communities of practice associated with value chain dimensions:** Importantly, each of the different value chain dimensions outlined above has associated **communities of practice**, i.e. groups of people that share a common value chain dimension, e.g. data scientists, IT specialists, developers of pre-trained models etc. These communities are part of the wider community of → **AI actors & communities**. While not always directly involved in AI, their work is an essential prerequisite for realising AI and preserving the values created.

**Figure 24.** Value chain dimensions. Value chain enablers in light green. Value chain assets in dark green.

| Value chain: assets and enablers   |                       |      |   |            |                                   |
|--|-----------------------|------|---|------------|-----------------------------------|
| Enabling IT infrastructure   | Enabling technologies | Data | Models<br>(pretrained, custom-made, foundation) | AI systems | Services supporting AI system use |
| Value-preserving enabler: Cybersecurity  |                       |      |   |            |                                   |
| Source: own production   |                       |      |   |            |                                   |
| <b>Explanatory note</b>  |                       |      |   |            |                                   |
| *) "An AI value chain is typically comprised of the following elements: computer hardware → cloud platforms → data and AI models → applications → services." ( <a href="#">UN, 2024a</a> ) |                       |      |   |            |                                   |
| **) in case of the use of pretrained models or foundation models as a basis for developing an AI solution or application case, such models can be considered enablers.                     |                       |      |   |            |                                   |
| <b>Term relationship:</b>  |                       |      |   |            |                                   |
| Related term   |                       |      |   |            |                                   |
| <ul style="list-style-type: none"><li>• Life cycle of AI in health</li><li>• AI value chain actors</li><li>• AI actors</li></ul>   |                       |      |   |            |                                   |

- AI practitioners

## Health technology assessment

**Cluster:** B.6 Evidence pathway for health

### Concept description

Health technology assessment (HTA) refers to the process of the evaluation of specific health technologies ([INAHTA, 2020](#); [European Commission, 2024d](#); [WHO, 2024c](#); [O'Rourke et al., 2020](#)), including medicinal products, medical devices ([Ming et al., 2022](#)) and prevention methods. HTA assesses technologies in regard to various aspects (e.g. medical, economic, social, ethics). Where necessary evaluations may draw on comparative analyses of the relative merit of products, e.g. superiority, inferiority, equality in regard to effectiveness and safety (for an example, see: [Park, 2023](#)). HTA is conducted by → health technology assessment (HTA) experts.

HTA ultimately aims at producing objective evidence for reimbursement decisions of healthcare technologies by paying organisations (e.g. health system organisations). AI introduced many novel complexities in the field of health products ([Alami et al., 2020](#)) and hence health technology assessment. These include a complex → value chain of AI and the resulting distributed responsibilities of data providers, providers of (pre-trained) models, enabling technologies and complex modelling decisions.

Given the potential important contribution of AI systems to medicine and healthcare (e.g. → AI-enabled medical device software) and the fact that AI systems in healthcare are considered health technologies, their real-world value for medical purposes needs to be assessed by HTA. To enable comparative assessment, HTA evaluators will require information on a variety of aspects including safety (→ AI safety; → NON-MALEFICENCE), → clinical performance and the claimed benefits of the AI system (→ BENEFICENCE) which should be supported by objective evidence from → clinical investigations, → clinical validation and → clinical evaluation.

We propose to consider HTA as a stage of the life cycle of AI systems in healthcare: AI developers are strongly encouraged to consider, already during design phases, how to generate *sufficient evidence* on product safety, on technological performance (typically in *research settings*), on ethical principles and on → real-world benefits (→ BENEFICENCE). This will support later HTA assessments and play a critical role for market success of health technologies.

Equally, HTA experts (→ Health technology assessment (HTA) experts) and clinical user communities (→ AI practitioners) should interact with AI developers in order to guide AI design, define knowledge gaps, gather experience (→ failure transparency) from the post-deployment / post-market space and elaborate information requirements in particular for clinical investigations (see [Farah et al., 2023](#)).

### Explanatory note

- The International Network of Agencies for Health Technology Assessment (INAHTA) has provided an updated definition of HTA ([INAHTA, 2020](#)): “*A multidisciplinary process that uses explicit methods to determine the value of a health technology at different points in its lifecycle. The purpose is to inform decision-making in order to promote an equitable, efficient, and high-quality health system.*” The definition is followed by four technical notes (see <https://htaglossary.net/health+technology+assessment>)
- A short description of HTA is provided on the EU Commission website on HTA ([European Commission, 2024d](#)): “*Health Technology Assessment (HTA) summarises information about medical, economic, social and ethical issues related to the use of a health technology. Examples of health technologies include medicinal products, medical equipment for diagnostic and treatment, prevention methods.*”
- The [WHO \(2024c\)](#) describes HTA as “*a systematic and multidisciplinary evaluation of the properties of health technologies and interventions covering both their direct and indirect*

*consequences. It is a multidisciplinary process that aims to determine the value of a health technology and to inform guidance on how these technologies can be used in health systems around the world. HTA is a transparent and accountable process that can be used by decision makers and other stakeholders to support the decision-making process in health care at the policy level by providing evidence about given technologies. It has been described as a bridge that connects the world of research to that of policy making."*

- [Ming et al., \(2022\)](#) provide an overview of and make recommendations for facilitating HTA for medical devices.

#### Term relationship:

Related terms:

- Clinical validation
- Clinical evaluation
- Clinical investigation

## Post-deployment monitoring

#### Cluster: B.6 Evidence pathway for health

#### Concept description

Post-deployment monitoring is a generic term that refers to the continuous monitoring of AI systems once they have been deployed (see for instance [Lovelace Institute, 2024](#)), i.e. made available to the intended users.

Notably, post-deployment monitoring may be regulated by legislations: for AI systems within the EU 'post-market monitoring' needs to be carried out under the AI Act ([EU, 2024a](#)). For AI systems covered under the EU's medical devices and in vitro diagnostic medical devices Regulations ([EU, 2017a, b](#)), relevant elements stipulated in the EU Act may be integrated, where applicable, into systems and plans already established in respect to these legislations (see explanatory note).

Information gathered during post-deployment monitoring may contribute to *inter alia*

- Assurance that relevant requirements are met.
- Understanding of performance, safety, → biases and any problems in the post-deployment space (e.g. → incidents, → adverse events) when using the AI system under real-world conditions. This will allow improvement and fixes including necessary communication on root causes (see → correcting problems and failures, including necessary communication; → failure transparency).
- Understanding of usability (see also → usability validation) as well as user concerns and proposals for usability improvements. This requires communication lines for user feedback (e.g. for AI tools used in health research and shared among research communities, feedback loops are essential to identify pitfalls and fix root causes).

Depending on the nature of the AI system, information gathered during post-deployment monitoring may contribute to a scientific understanding and broader debate of wider impacts, e.g. on behavioural aspects of users, impact on skills, the workplace, education and learning, especially of next generations or impacts on specific persons, groups, communities and the society as a whole. In health this may concern, discussions concerning → AI and the development of healthcare.

For AI systems used in health research, public health, health system management), post-deployment monitoring may also entail ongoing → peer review and community discourse.

## Explanatory note

Notably, in case of health products (e.g. AI-enabled medical devices), the well-established concept of post-market surveillance ([IMDR, 2006-2012](#)) covers relevant legislative and regulatory needs in relation to quality, safety and performance (see → [post-market surveillance, market surveillance, corrective action](#)). Post-market surveillance activities should be “*proportionate to the risk class and to the type of device*” ([EU, 2017a](#); Article 83).

Similarly, the EU's AI Act requires post-market monitoring activities that are “*proportionate to the nature of the AI technologies and the risks of the high-risk AI system*” ([EU, 2024a](#); Article 72). The AI act also stipulates that for high-risk AI systems covered by other EU legislation (e.g. medical devices), providers have a choice of integrating the AI Act's requirements (Article 72) into systems and plans already existing under that relevant legislation (e.g. relating to post-market surveillance of medical devices), provided that an equivalent level of protection is achieved.

## Term relationship:

Related terms:

- [Post-market surveillance](#) (for medical technology including AI systems used for medical purposes. The term is used by various jurisdictions; see IMDRF / GHTF study group ‘post market surveillance/vigilance’)
- Post-market monitoring (see EU's AI Act, Article 72)

## Decommissioning / Retirement

### Cluster: B.6 Evidence pathway for health

#### Concept description

Decommissioning refers to the intentional cessation of operation of an AI system, e.g. because it reached the end of its life cycle, is not any more updatable or obsolete for various reasons (e.g. due to → [drift / shift in machine learning](#)). Decommissioning may involve the deployer, the developer and the user (→ [users of AI in health and medicine](#)) as well as other relevant actors.

- Notably, “*decommissioning*” does not include circumstances of cessation of operation covered by regulatory terms such as “*withdrawal*” or “*recall*” (e.g. EU's AI Act; [EU, 2024a](#)).

Ideally, decommissioning should be conducted as a consistent process. However, practical aspects of decommissioning will be influenced by the → [use environment](#) and → [use context](#) of the AI system in question.

Ideally, a decommissioning plan should be developed already at early stages of the → [life cycle of AI in health](#), to ensure effective decommissioning when the need arises.

#### Explanatory note

Decommissioning should address various aspects, including

1. [Planning decommissioning](#): Includes a mapping of dependencies and relationships of the AI system with other (local) assets, i.e. IT infrastructure and other enabling technologies ([Reina & Griesinger, 2024](#)). These include local storage, cloud storage, networks, and other systems. This allows anticipating the impacts on the → [use environment](#) and plan the decommissioning process accordingly. In particular, the impact on the → [use environment](#)'s business continuity and users' and stakeholders' capacity to perform their work once the

- system has been decommissioned needs to be analysed beforehand, including timely planning for alternative solutions.
2. Documentation and archiving: a) of the decommissioning process covering the points below, b) where possible models/code, documentation with relevance for potential audits or in relation to legal compliance.
  3. Data management: issues of data retention, data deletion (secure destruction erasure from physical media, deletion from cloud services according to service provider's specifications), data archiving, migration of data to other platforms (e.g. to integrate into existing research databases) and, importantly, data privacy (e.g. in line with EU's GDPR ([EU\\_2016b](#)); see also 'legal compliance' below) concerning → **post-deployment input data**. This may include data of patients, (logging) data of healthcare professionals/users etc.  
Concerning secure data deletion, caution must be taken not to simply delete files (which removes these from directories while leaving the data unchanged) but to follow secure methods for destroying / erasing data files completely, e.g. using overwriting software (see for instance NIST standard 800-88; [NIST, 2014](#)).
  4. Contractual obligations with commercial or non-commercial partners, e.g. maintenance contracts, research consortia, cloud service provider etc.
  5. Legal compliance: ensuring that the decommissioning respects requirements of relevant and applicable legal frameworks (e.g. EU's GDPR in regard to data privacy; [EU\\_2016b](#))
  6. Alignment with existing processes (e.g. quality assurance): depending on use environment, the AI system may have operated in the context of specific processes and (locally applicable) management systems, which need to be respected during decommissioning.
  7. Communication: ensuring that all relevant parties receive information about the fact that the AI system has been decommissioned and are alerted to alternative solutions.
  8. Impact on use environment and business continuity: minimisation of the decommissioning's impact on the → **use environment** in terms of business continuity. Identification and installation of other systems or solutions prior to decommissioning.
  9. (Cyber) security: in case a decommissioned AI system is not erased and completely shut down, network connections (either physical or through wireless means) are disabled to prevent unauthorised access to remaining code/model and relevant data that had not been deleted/ erased (see point 3 on Data management).

Amazon has published a best practice guide for retiring applications, which is useful for a wider range of digital tools ([Amazon web services, 2024](#)).

#### **Term relationship:**

Synonyms:

- Retirement  
(see for instance the OECD report ([2024a](#)) on the implementation of the OECD recommendation on artificial intelligence, which uses "retirement" and "decommissioning" interchangeably).

Related terms:

- Asset decommissioning
- Data destruction, data erasure
- Withdrawal
- Recall

## B.7 Data

### Dataset

**Cluster:** B.7 Data

#### Concept description

A collection of data. Datasets may be mono-modal, i.e. composed of data of only one → data modality or → multi-modal, composed of data of more than one modality.

### Development data

**Cluster:** B.7 Data

#### Concept description

Data for developing an AI model, typically consisting of → training data, → validation data and → testing data, compiled in three datasets which are often achieved by partitioning one single dataset into these three subsets.

#### Term relationship:

Related terms:

- Training data
- Validation data
- Testing data

### Data provenance of development data

**Cluster:** B.7 Data

#### Concept description

Data provenance refers to origin of the → development data, i.e. the data source, whether collected by the developer or acquired through a third-party data provider, whether synthetic (→ synthetic health data), whether directly observed or somehow derived, the level of aggregation and data processing (→ data processing / wrangling).

#### Explanatory note

Definition based on the OECD's framework for classification of AI systems (OECD, 2022).

## Training data

### Cluster: B.7 Data

#### Concept description

We use the definition of Liu et al., (2022): "A dataset of instances used for learning parameters of a model".

In supervised learning (→ AI technique) the input data and associated labelled output data form the (labelled) training data.

#### Term relationship:

Related terms:

- Parameters
- Hyperparameters
- Dataset
- Development data
- Validation data
- Testing data

## Validation data

### Cluster: B.7 Data

#### Concept description

The set of data used for → model validation. Validation data are distinct from → training data and → testing data, although all three sets are usually obtained by slicing up one initial data set into these three. Validation data serve to obtain more information on the initial 'validity' of the model, i.e. its performance under laboratory conditions for the intended purpose. Validation data support the fine tuning of models (the validation data set is sometimes also referred to as 'tuning set' (Park et al., 2021), e.g. for adjusting learning rates (→ hyperparameters\*), modification of network architecture or regularisation to reduce → overfitting).

\*) validation data can be seen as "A dataset of instances used to tune the hyperparameters of a model" Liu et al., (2022).

#### Term relationship:

Related terms:

- Dataset
- Development data
- Training data
- Testing data
- Model validation

Synonyms:

- Tuning set (i.e. tuning data set)

## References and further reading

- See → training data

## Testing data

### Cluster: B.7 Data

#### Concept description

The set of data used for → model testing. The testing data set must be completely separate from the → training data and → validation data. However, the testing data should follow the same distribution as the training data ([Liu et al., 2022](#)) so as to allow an assessment of the performance of the model as intended and trained, but with additional insights into potential model → generalisability.

#### Term relationship:

Related terms:

- Dataset
- Development data
- Training data
- Testing data
- Model testing
- Model evaluation

Synonyms:

- Holdout dataset (in case part of the original dataset is set aside as a testing dataset) ([Liu et al., 2022](#)).

#### References and further reading

- See → training data

## Input data

### Cluster: B.7 Data

#### Concept description

Data provided to an AI system or directly acquired (e.g. via sensors) by an AI system or a model on the basis of which the system or model produces an output (e.g. recommendation, prediction, action, content). *Description based on definition 33 of the EU's AI Act ([EU, 2024a](#))*.

During → machine learning, the → training data are the input data on which a → machine learning algorithm is run to create a → machine learning model.

#### Explanatory note

Input data can be → development data (i.e. → training data, → validation data, → testing data) or → post-deployment input data. Notably → post-deployment input data can also be used for training an AI system. This holds for both non-continuously learning AI systems and continuously learning AI systems.

For the latter ones, continuous training with → post-deployment input data may happen in the post-deployment phase. For non-continuously learning AI systems, training with → post-deployment input data would most likely occur during renewed development. Input data will have one or more → data modality.

## Term relationship:

Related terms:

- Development data
- Post-deployment input data
- Sensors

## Output and output data

### Cluster: B.7 Data

#### Concept description

All output generated by an AI system (→ AI system output) can be understood as data in itself which may be used for assessing, evaluating, adjusting, improving an AI system or, inversely, leading to its decommissioning (→ decommissioning / retirement).

Secondly, the term “output data” is increasingly used to denote specifically content synthesised (or “created”) by → foundation models; generative AI, based on → input data and the AI system’s → algorithm or → model.

#### Explanatory note

AI-generated data (that is output data of generative AI) that are increasingly available on internet servers are recognised as a risk for the development and training of future AI models (Veselowsky et al., 2023): internet-derived data are often used for training → foundation models; generative AI, leading to a vicious circle of data and model degradation: training such models with AI-generated data (in particular “hallucinated” or “confabulated” data) may “poison” subsequent or new versions of AI systems, i.e. train them on false, fabricated, misleading data (Anabheri, 2024).

This poses an additional risk in addition to risks originating from unexamined biases that may exist in historical data sets (e.g. Pubmed data base), such as outdated scientific concepts (e.g. biomarkers) or deeply encapsulated racial biases. AI systems, if uncritically adopted, may simply reflect “a repository of the collective medical mind” (Char et al., 2018).

## Term relationship:

Related terms:

- Input data
- Development data
- Post-deployment input data

## Data quality

### Cluster: B.7 Data

#### Concept description

Intuitively, the quality of data is essential for their successful use and exploitation for various purposes, notably for analysing and using health data in large data sets for health research (including through AI) and using data for training → machine learning models and AI systems. What precisely constitutes data of ‘high’ quality has been debated extensively, leading to standardised approaches and methods for quality assessment (→ data quality metrics) to systematise the consistent description, measurement

and documentation of data quality. However, given the fluid nature of data, the emerging development of big health data, data quality will likely remain a topic that requires constant reassessment.

### **Challenges of data quality**

There are several challenges of data quality, in particular, when dealing with big data (e.g. large health data sets) ([Cai & Zhu, 2015](#)):

- A high diversity of data sources that results in various (complex) data structures. This makes data integration difficult.
- A high data volume complicates the efficient assessment of data quality.
- Some data change fast and, consequentially, so-called “timeliness” of data ([McGilvray, 2010](#)) may be short-lived. Depending on application, continuous collection and fast processing of data may be required.

For machine learning and AI, data quality of the → development data has a decisive impact on various aspects of an → machine learning model, including

- the overall quality of its → output and output data,
- the model’s performance (e.g. → accuracy)
- the model’s robustness (→ risks related to insufficient robustness / resilience), → reliability, → generalisability, → FAIRNESS,
- the level of → AI safety exhibited by the model / AI system
- the model’s scientific and/or clinical relevance (→ conceptual relevance, → valid clinical association / scientific validity) and interpretability and/or explainability (→ interpretability and explainability).

### **Explanatory note**

#### **Examples of data quality issues**

Data quality issues include for instance incomplete data (which, depending on the type of incompleteness, could lead to → bias), incorrectly formatted or coded data, false or truncated data or data that contain duplicates (which misleadingly suggests data richness and could lead to → overfitting), data that are not readily accessible, not credible, inaccurate or irrelevant in regard to the → conceptual relevance and/or → contextual relevance.

**N.B. Data quality issues should not be confused with → data drift / shift.**

#### **Data quality dimensions, elements or characteristics**

Data quality has dimensions that are generally applicable (e.g. correctness of data) and others that are relational, such as fitness-for-purpose of the data for a given modelling project relating to a real-world problem (→ model development).

Four basic data quality dimensions have been proposed by [Wang & Strong, 1996](#):

- intrinsic
- contextual (what we refer to as ‘relational’)
- representational
- accessibility

[ISO / ICE \(2008\)](#) has proposed in their standard 25012 a data quality model composed of 15 data quality characteristics that relate to aspects of inherent data quality (degree to which data quality characteristics may satisfy needs when data are used under specified conditions) and system-dependent data quality (degree to which data quality is maintained within a computer system when data is used under specified conditions).

[Cai & Zhu \(2015\)](#) have proposed detailed data quality assessment criteria specifically for big data, using a system of 'dimensions' and 'elements' that partly overlap with ISO standard 25012, but are expounded in view of big data, i.e. the 'dimensions' of availability, usability, reliability, relevance and presentation quality) and the 'elements' of accessibility, timeliness, credibility, accuracy, consistency, integrity, completeness, fitness, readability.

[Black & Van Nederpelt \(2020\)](#) have provided a detailed analysis of data quality dimensions, compiling definitions of relevant associated terms from various sources.

[Bernardi et al. \(2023\)](#) provide a comprehensive review on data quality in health research.

[Zhou et al. \(2024\)](#) have surveyed data quality dimensions and tools.

#### Term relationship:

Related terms:

- [Data quality metrics](#)
- [Development data](#)
- [Training data](#)
- [Input data](#)

## Data quality metrics

### Cluster: B.7 Data

#### Concept description

Data quality metrics in the context of AI refers to the consistent or even standardised manner of describing, measuring and documenting various aspects of input data that are used as → [development data](#) (in particular → [training data](#)). There is currently no clear consensus on quality domains or data quality metrics and there is no generally accepted consensus frameworks. Proposals include the standard on 'measurement of data quality' by ISO / IEC's ([ISO / IEC, 2015a](#)) and the data assessment framework by [Cai & Zhu \(2015\)](#) and [Elouataoui et al. \(2022\)](#) regarding big data.

[Ehrlinger & Wöß \(2022\)](#) have published a survey of data quality measurement and monitoring tools. [Schwabe et al. \(2024\)](#) have examined data quality frameworks and combined the results with considerations of data quality in medicine, resulting in a framework of 15 awareness dimensions.

This highlights the need for collaborative work and/or standardisation activities.

We propose that data quality metrics for AI applications in health and medicine should address I) *general data quality aspects* (e.g. correctness, no duplications) and II) *relational data quality aspects*, i.e. whether the given data are fit-for-purpose for the development of a model / AI system with a specific → [output](#) and [output data](#), → [intended use](#) and - > [use context](#) (most technical aspects in the explanatory note below are "relational").

In addition, since data for AI applications in health are typically based on private and sensitive data, quality metrics should address also aspects of how stored (health) data are managed in terms of data protection and data security (→ [PRIVACY PROTECTION](#)).

#### Explanatory note

Data metrics should help ensuring that high-quality data are used for training a → [learning algorithm](#) so as to ensure that the final → [model](#) of the AI system produces relevant (→ [conceptual relevance](#); → [valid clinical association / scientific validity](#)) and accurate (→ [accuracy](#)) outputs in relation to the → [intended use](#).

To represent principles of trustworthy AI in data quality aspects, we propose that data quality metrics should cover the following elements, based on considerations in Griesinger et al. (2022); ISO /IEC (2015a), Schwabe et al., (2024):

- Data quality metrics will have to cover to some extent ethical principles and concepts such as
  - Absence of bias ( $\rightarrow$  NON-MALEFICENCE)
  - Fairness, inclusion, non-discrimination ( $\rightarrow$  FAIRNESS)
  - Patient consent regarding collection, processing and use of health data ( $\rightarrow$  DIGNITY, FREEDOM AND AUTONOMY), unless  $\rightarrow$  synthetic health data or highly aggregated data and/or “de-identified” (e.g. pseudoanonymised) data are being used.
  - Given the sensitivity of health data, data quality metrics should also measure to which extent data that are being (iteratively) used for training a  $\rightarrow$  machine learning algorithm are protected, e.g. to avoid unauthorised access to data ( $\rightarrow$  PRIVACY PROTECTION) as well as provisions to allow access of subjects to their health data ( $\rightarrow$  PRIVACY PROTECTION).
- Technical aspects of data quality metrics should include:
  - Completeness of dataset, i.e. are the data complete or are there missing data points or larger gaps (incompleteness could lead to representation bias ( $\rightarrow$  bias)). Is the labelling complete?
  - Correctness of data, i.e. are the data correct, devoid of errors, not truncated, devoid of anomalies and implausible outliers, devoid of duplicates (which misleadingly suggests data richness and could lead to  $\rightarrow$  overfitting) and any other elements of confusion or disorganisation?
  - Currentness of data, i.e. are the data up to date to be sufficiently fit-for-purpose for a given  $\rightarrow$  intended use and/or  $\rightarrow$  use context in particular considering potential  $\rightarrow$  concept drift / shift.
  - Symmetry, skewedness, kurtosis of dataset / data distribution (kurtosis of 0 = normal distribution; in negatively skewed data mean is before median is before mode; in positively skewed data mode is before median is before mean).
  - Inter- and intra-data set consistency, i.e. is there consistency of key descriptors/parameters,  $\rightarrow$  labels / data labels,  $\rightarrow$  features,  $\rightarrow$  attributes, especially where datasets have been derived from various sources (e.g. under supervision of different data scientists) and/or aggregating in an automated manner.
  - Representativeness of the real-world problem (see also  $\rightarrow$  bias). Depending on the  $\rightarrow$  intended use of the AI system, representative may relate to highly inclusive data sets (e.g. regarding sensitive attributes: universal design / application) or to specifically targeted datasets (e.g. targeted design for specific patient groups, vulnerable groups, diseased patients etc.).
  - Balancedness (e.g. in view of achieving statistically meaningful predictive outcomes)
  - Drift: are the data up to date in view of  $\rightarrow$  conceptual relevance and are the data sufficiently current (see currentness above) so that  $\rightarrow$  data drifts / shifts can be reasonably excluded at least for the mid-term future of using the AI system?
  - Appropriateness of selection of  $\rightarrow$  features,  $\rightarrow$  attributes,  $\rightarrow$  proxies, including an assessment of their relevance ( $\rightarrow$  conceptual relevance;  $\rightarrow$  valid clinical association / scientific validity) for the intended  $\rightarrow$  output and output data and  $\rightarrow$  intended use.

#### Term relationship:

Related terms:

- Data quality
- Development data
- Training data

- Input data

## Post-deployment input data

**Cluster:** B.7 Data

### Concept description

Post-deployment input data are → input data during the post-deployment phase. These data may be acquired (e.g. radiological imaging) and fed into an AI system or acquired directly by an AI system (e.g. fitted with sensors for a specific data modality).

### Term relationship:

Related terms:

- Input data

## Data modality

**Cluster:** B.7 Data

### Concept description

Data are the “information stream” of artificial systems, including AI. In analogy to the neurobiological term “sensory modality”, we propose to categorise data according to their modality, e.g. visual, auditory, electromagnetic, textual etc. We propose the following data modalities:

- **Image data**, including other modalities transformed to images (e.g. a power spectral density graph of a sound signal, x-ray data converted into an image, acoustic data, e.g. ultrasound, converted to an image)
- **Acoustic or ‘audio’ data**: sound pressure data, including spoken language or ‘speech’
- **Textual data** based on symbols with semantic meaning for humans (written human language, including computer-readable code, genomics data)
- **Electromagnetic field data** (e.g. radar, x-rays)
- **Mechanical / tactile data** (e.g. contact/touch, surface scanning, vibration)
- **Data on chemical structures and their interaction with other molecules**, including biological ligands/receptors, epitopes
- **Knowledge base** (may integrate various data modalities)
- **Logical data** (e.g. ontological relationships, causal relationships, Bayes nets etc.)
- **Metadata** (to describe data modalities).

### Explanatory note

Data modality will in practice overlap to some extent with various labels of → AI typology. For instance, AI processing image data can also be termed “visual AI” etc.

### Term relationship:

Related term:

- AI typology

## Nature of data

### Cluster: B.7 Data

#### Concept description

Refers to whether the data are dynamic, static, dynamic updated from time to time or real-time.

#### Explanatory note

Concept description based on the OECD classification framework of AI systems ([OECD, 2022](#)).

#### Term relationship:

Related term:

- [Data modality](#)

## Datasheets for datasets

### Cluster: B.7 Data

#### Concept description

Datasheets for datasets have been proposed as a way of addressing the issues that there are, currently, no standardised processes and/or templates for documenting datasets within the machine learning community. Dataset should be *“accompanied with a datasheet that documents its motivation, composition, collection process, recommended uses, and so on. Datasheets for datasets will facilitate better communication between dataset creators and dataset consumers, and encourage the machine learning community to prioritize transparency and accountability.”* ([Gebru et al., 2021](#)).

#### Explanatory note

This proposal should be seen in the context of efforts → [data quality metrics](#) and may constitute an important component towards more transparency of datasets, even where these are not physically available, e.g. in instances where models were trained on sets of health data distributed in different countries and which could not be transferred (e.g. for privacy or legal reasons) to the AI developer.

The STANDING Together initiative promises to develop documentation recommendations for data used for training AI models in health and medicine ([Ganapathi et al., 2022](#)).

#### Term relationship:

Related terms:

- [Data quality metrics](#)

## Synthetic health data

### Cluster: B.7 Data

#### Concept description

Synthetic health data are all health-relevant data that have been created through algorithms or models in view of addressing a specific data-dependent aim (scientific, clinical etc.). Synthetic data typically recapitulate or reflect relevant features and statistical properties of real-world data without however

making direct use of these. Baowaly et al. (2019), Giuffrè & Shung (2023), Ayilara et al. (2023) demonstrate the generation and use of synthetic health data. Characterisation and evaluation of synthetic health data see Lenatti et al., (2023) and Abdusalomov et al., (2023).

Synthetic data hold the promise to address the difficulty of accessing health data for training and developing AI systems due to obstacles related to data privacy, patient consent and/or lack of availability of electronic health data. However, synthetic may also pose novel legal and policy challenges (European Commission, 2024e).

### Explanatory note

There is no consensus yet on the precise meaning of “synthetic data”. A proposal put forward by Jordan et al. (2022) defines synthetic data as *“data that has been generated using a purpose-built mathematical model or algorithm, with the aim of solving a (set of) data science task(s)”*.

However, synthetic data are often understood as newly created data that mirror the statistical characteristics of an original data set, where relevant patterns and relationships in, for example, a real-world administrative health data set are preserved (see for instance: Mosquera et al., 2023).

The main types of synthetic data for clinical purpose include tabular, time-series, text-based data, synthetic images or videos as well as audio content (Giuffrè et al., 2023).

### Term relationship:

Related terms:

- Development data
- Training data
- Input data
- Data privacy

## Data privacy

### Cluster: B.7 Data

#### Concept description

Data privacy means that persons should have control over their → personal data, i.e. who collects, stores, processes these and how they are used. Data privacy has many aspects, but it is particularly relevant in the area of digital technologies and AI, which depend on data. The development of AI systems for healthcare and health research requires personal health data for training a → machine learning model, for → model validation and testing (see → training data, → validation data, → testing data).

What concerns AI systems in health, data privacy is a key concern (Ahmed et al., 2023; Brady & Neri, 2020; Kelly et al., 2019). Data privacy in general and in the context of health is rooted in relevant international conventions; for details, see → medical privacy / health privacy. Patients must be asked for their consent in case their data are collected, stored, processed and used (→ consent concerning collection of personal data / medical information). Patients must be informed how their data are being used (i.e. for what purpose), including in case their data are shared with AI developers (e.g. local hospital research teams or companies) for purposes of training/developing new AI models (e.g. for health research, healthcare). Patients should be made aware about the eventual fate of their data, including protocols for data destruction, including in case of the → decommissioning / retirement of AI system's under evaluation (e.g. in the context of → model evaluation and post-processing exercises or → clinical investigation prior to conformity assessment and/or authorization. Patients need to know their legal access rights to their data and what steps are taken to ensure privacy protection and → data security.

Data privacy is protected by legislative frameworks, in the EU the General Data Protection Regulation, GDPR (EU, 2016b). Notably, the EU's Data Act (EU, 2022) does not regulate the protection and privacy

of personal data. Instead it provides rules for data sharing, access and use in order to enhance the distribution and availability of data within the EU's data economy. This has relevance for data within the → value chain of AI. The EU's AI Act ([EU, 2024a](#)) refers to the GDPR and other relevant legislations in the context of data and data governance (see EU AI Act, Article 10).

#### Explanatory note

**See also → PRIVACY PROTECTION** and translational concepts.

Article 5 of the EU's GDPR ([EU, 2016](#)) stipulates the principles relating to processing of personal data. Additionally, Articles from 12 to 23 in Chapter 3 of the EU's GDPR ([EU, 2016](#)) list and present the rights of the data subject. These articles are grouped in five sections:

- Transparency and modalities
- Information and access to personal data
- Rectification and erasure
- Right to object and automated individual decision-making
- Restrictions

Recital 10 of the EU's AI Act ([EU, 2024a](#)) clarifies that this Regulation does not intend to modify the current application of Union law (i.e. EU's GDPR; [EU, 2016](#)) regarding the processing of personal data. Article 10 of the EU's AI Act stipulates that high-risk AI systems must be developed using data sets that should be managed effectively and should be relevant and representative. Furthermore, when processing special categories of personal data, data providers must adhere to strict conditions to safeguard individuals' rights and freedoms.

#### Term relationship:

Related terms:

- **DIGNITY, FREEDOM AND AUTONOMY**

## Personal data

### Cluster: B.7 Data

#### Concept description

We refer to the definition of the EU's GDPR ([EU, 2016b](#)):

*'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person'.*

#### Explanatory note

Article 4 of the EU GDPR provides also definitions of other relevant terms in this context.

#### Term relationship:

Related terms:

- **DIGNITY, FREEDOM AND AUTONOMY**

## CIA principles

### Cluster: B.7 Data

#### Concept description

The CIA principles, also known as the 'CIA Triad' ([Unitrends, 2024](#)), form a fundamental model in information and cyber security that consists of three core components: Confidentiality, Integrity, and Availability. The CIA triad serves as a foundational guideline for organisations, system administrators and developers to design and implement effective cybersecurity policies and practices ([Osaro, 2023; Lundgren & Möller, 2017; NIST, 2017a](#)).

##### *Confidentiality:*

Confidentiality, identifies the "*property that information is not made available or disclosed to unauthorized individuals, entities, or processes*" ([ISO/IEC 2018](#)). This property is satisfied when, for example, sensitive information is only accessed by authorized users or systems.

##### *Integrity:*

Integrity, which may include authenticity and non-repudiation, identifies the "property that data has not been altered or destroyed in an unauthorized manner" ([ISO/IEC 2018](#)). Data integrity is satisfied when data remains accurate, complete, and unchanged without unauthorized modifications whereas system integrity is maintained when it operates as intended, without any unauthorized interference that could compromise its performance or functionality.

##### *Availability:*

Availability, identifies the "property [of data and services] of being accessible and usable upon demand by an authorized entity" ([ISO/IEC 2018](#)). This property is satisfied when authorised users or processes can access the required data and can operate the requested services.

#### Explanatory note

The CIA triad is often shown as a triangle, with each vertex representing one of the three principles: confidentiality, integrity, and availability.

These three principles are connected, meaning that if one is compromised, it may impact the others. For example, if integrity is compromised, availability could be at risk.

Organisations can use various security controls to protect these the CIA principles:

- Technical controls such as firewalls, intrusion detection systems, and encryption.
- Administrative controls such as user training, security policies and procedures.
- Physical controls such as access control measures and perimeter security.

By understanding these three goals and applying the right security measures, organisations can protect their information and reduce the risks of cyber security threats ([ENISA, 2016; Reina & Griesinger, 2024](#);

#### Term relationship:

Related terms:

- Training data
- Validation data
- Testing data

## Sample / Sampling

### Cluster: B.7 Data

#### Concept description

A sample is a subset of a larger set of data. Sampling is process of selecting this subset (sample) of data from the total available data, e.g. to estimate characteristics of the whole population ([Cowan, 2024](#)).

Samples that are deemed representative for the larger set and allow the more efficient training of AI systems with smaller data sets, thus reducing costs and environmental footprint of AI development.

Data sampling is evidently an important potential source of → bias, in case samples are not fully representative but skewed towards specific → features or → attributes. Sampling must not be confused with the use of → proxies.

#### Explanatory note

For an introduction to methods and types of sampling, see Cowan ([2016](#)) and Chirag ([2023](#)).

#### Term relationship:

Related terms:

- Dataset
- Development data
- Training data
- Data processing / wrangling
- Feature
- Attributes
- Proxies
- Bias

## Data processing / wrangling

### Cluster: B.7 Data

#### Concept description

Steps taken after creation and/or acquisition of datasets in order to address e.g. missing values or irregular sample data. These steps should be properly recorded (→ traceability) since they may contribute to → bias that is subsequently, during training the model, integrated into the model ([Liu et al., 2022](#)).

#### Explanatory note

N.A.

#### Term relationship:

Related terms:

- Bias

## Data FAIRification

### Cluster: B.7 Data

#### Concept description

Measures to ensure that data are Findable, Accessible, Interoperable, and Reusable = ‘FAIR’ so as to overcome data management challenges including “service incompatibilities, data access restrictions, unavailability of data, missing data, and incomplete, ambiguous or absent metadata.” ([Welter et al., 2023](#)). A key publication on the principles of data FAIRification is “The FAIR guiding principles for scientific data management and stewardship” ([Wilkinson et al., 2016](#)).

#### Term relationship:

Related terms:

- Data management

## Attributes

### Cluster: B.7 Data

#### Concept description

Generally, an attribute is a quality, character, or characteristic ascribed to someone or something. In line with this general meaning, ‘attribute’ is considered a synonym of ‘feature’ in the context of data science and machine learning ([Google, 2024](#)).

However, attribute or *sensitive attribute* is often used in the context of considerations of → FAIRNESS in machine learning, referring typically to characteristics or features of a person, sub-population or group which may be correlated with consequential decision making resulting in indirect discrimination (e.g. [Datta et al., 2017](#); [Haeri & Zweig, 2020](#); [Shah et al., 2023](#)).

#### Explanatory note

In the context of healthcare (and other fields of application), attributes such as race, gender, sex, ethnicity, skin colour, national origin, sexual orientation are *sensitive* attributes. Specific care must be taken to avoid possible discrimination (→ FAIRNESS) or → bias that might occur due to various sources, including: non-critical use of historically biased data, training data that have not been carefully compiled or processed concerning bias assessment and reduction, conceptual bias when taking modelling decisions, including through the choice of → proxies for decision making or for identifying and discriminating → sample / sampling in a fair manner.

#### Term relationship:

Synonyms:

- Feature

Related terms:

- Data processing / wrangling
- Labels / data labels
- Features
- Proxies
- Development data
- Training data

## Features

### Cluster: B.7 Data

#### Concept description

Features are properties or characteristics of data relating to a specific real-world phenomenon (see [Bishop, 2006](#)). In a more restricted sense, features are input variables for a machine learning model ([Google ML glossary](#)). Features must be measurable or machine-describable to serve as input variables to a → machine learning algorithm for developing a → machine learning model.

Selecting relevant features is a key element for building models that show appropriate performance (e.g. for classification, regression, pattern recognition etc.) in regard to the real-world phenomenon or problem to be modelled (→ conceptual relevance).

For AI in medicine and healthcare, appropriate features need to be chosen that adequately represent original health data: features need to relate in a meaningful and valid way to the problem to be tackled, e.g. predicting lung cancer from radiological images or predicting disease-risk based on appropriate clinical, socioeconomic and other features ([Roski et al., 2022; Liu et al., 2022](#)) (see → valid clinical association / scientific validity).

In case of supervised learning, features are used by a 'supervised model' to predict the → label / data labels, i.e. the outcome of the model. Typical outcomes include a classification prediction, a value or a recommended action (e.g. AI-enabled robotic surgery).

#### Example:

In a diagnostic model that predicts intraocular retinoblastoma (see for example [Kaliki et al., 2023](#)), the features (e.g. extracted from fundus images) could be intraocular haemorrhage, blood vessel morphology, presence of seeds (macrophages) etc. The label could be binary, i.e. 'positive' or 'negative' for retinoblastoma.

#### Explanatory note

Features are key elements of data sets. Data sets consist of individual 'examples'. Each example consists of one or more features and, in case of supervised learning, a → label / data labels.

For instance, when training a model to predict likelihood of coronary heart disease, the features '*body mass index*' and '*blood pressure*' may constitute the input variables required for producing the output relating to 'coronary heart disease', e.g. as a score or a prediction. In this case, individual feature combinations (e.g. from various patients) would be called 'examples'.

#### Term relationship:

Synonyms:

- Variables
- Input variables
- Attributes

Related terms:

- Data processing / wrangling
- Labels / data labels
- Attributes
- Development data
- Training data

## Labels / data labels

**Cluster:** B.7 Data

### Concept description

Labels are meaningful annotations or tags associated with examples in a data set (see → [features](#)). Associating data with labels is called "data labelling" or "data annotation" ([Liu et al., 2022](#)).

Labelling data is necessary for supervised machine learning, where the machine → [learning algorithm](#) is presented labelled data that allow the algorithm to optimise the chosen → [model](#) for the desired outcome (→ [output](#) and [output data](#)), e.g. prediction of a class = classification; prediction of a value.

Labels are essential for yielding AI systems with high levels of → [accuracy](#). Data labelling therefore is a key quality-defining step when processing collected data. In contrast to → [proxies](#) that are used as surrogates for data features, labels are simplified but direct annotations of examples and their features within a dataset.

### Term relationship:

Related terms:

- [Data processing / wrangling](#)
- [Features](#)
- [Attributes](#)
- [Development data](#)
- [Training data](#)

## Proxies

**Cluster:** B.7 Data

### Concept description

Generally, proxies serve as a surrogate or substitute for more complex properties, characteristic, variables, measures or, generally, phenomena observed in a real-world situation or problem. In statistics and machine learning "proxies", "proxy variables" or "surrogate variables" are measurable and/or quantifiable variables that cannot readily be measured ([Upton & Cook, 2014](#)).

In AI models, proxies can be used for predicting the behaviour of more complex systems. Proxies offer a number of advantages for modelling, but do also have downsides: The use of proxies has various benefits such as reducing computational complexity (thus enhancing computation efficiency and thus environmental → [SUSTAINABILITY](#)), allowing to deal with incomplete data or data with low signal to noise ratio. Proxies may render models also more interpretable since proxies simplify real-world problems, rendering the AI logic more accessible to human reasoning: proxies may allow to understand in a more immediate manner why a specific output (e.g. recommendation) was generated.

Downsides of proxies include conceptual oversimplification of complex systems, potentially lower → [accuracy](#) or reduced → [conceptual relevance](#) of AI model outputs (→ [outputs](#) and [output data](#)) Thus, proxies need to be chosen or designed with great care so as to balance modelling efficiency with output accuracy and relevance.

When used in decision tools, proxies that refer to legally protected → [features](#) or sensitive → [attributes](#) may lead to 'algorithmic proxy discrimination' ([Chen, 2024](#)). While 'proxy discrimination' is still debated in terms of precise meaning of the term and its relationship to statistical dependencies, causal effects, and intentions (e.g. [Tschantz, 2022](#)), there is widespread attention to the risk of propagating → [bias](#)

(e.g. 'label bias') and possible discrimination ([Datta et al., 2017](#)) - for instance in the context of risk assessments, including identification of high need patients ([Zanger-Tishler et al., 2024](#)). Fagan ([2025](#)) pointed out that there is no appropriate 'test' "*for regulating the use of variables that proxy for race and other protected classes and classifications*", proposing a comparative approach to tackle this seemingly technical issue that may have considerable impact on → FAIRNESS.

### Explanatory note

#### **Examples of proxies**

- Intensity of pixels at a specific topological environment or a specific location in a radiological image can be used as a proxy or surrogate for more complex object properties.
- Postcodes have been used as indicators or predictors of various properties, including as markers for socioeconomic status (poverty, class) that have been associated with risk of disease and potential causes of disease ([Danesh et al., 1999](#)) or disability ([Moola, 2024](#)).
- Medical concepts in publicly available knowledge sources have been used for extracting features that can be used as surrogates or proxies to 'gold-standard' data labelling in the context of high-throughput clinical phenotyping ([Yu et al., 2017](#)).

N.B. Proxies in AI/machine learning models for health applications should not be confused with clinical "surrogate endpoints" (see for instance: [Manyara et al., 2024](#)), despite being based on substituting more complex observables.

### Term relationship:

Synonym

- Proxy variables
- Surrogate variables

## B.8 Algorithm, model, algorithm-to-model transition

### Algorithm

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

#### Concept description

A process or set of mathematical instructions or rules laid down in computer-readable code and used by a computer for making calculations or executing complex (problem-solving) operations. All artificial intelligence solutions are based on algorithms. Some are 'classical' algorithms, produced without machine learning (→ non-learning algorithms). However, AI is nowadays dominated by → machine learning algorithms and their output, → machine learning models, both being obviously algorithms or programmes (Brownlee, 2020):

- A machine → machine learning algorithm is a procedure that is run on → training data and produces another program, the machine learning model.
- A → machine learning model processes real-world → input data; its specific program (the 'model', developed during machine learning) produces an output (→ output and output data) that is aligned with the → intended use of the AI system.

Output examples are classification predictions, detection of objects, segmentation into regions of interest (ROIs), predictions of continuous values, recommendations, executing a movement via actuators etc.

#### Explanatory note

AI employs a wide range of algorithms to realise AI models employed in AI systems aimed to achieve specific objectives or solve specific problems.

Examples of machine learning algorithms (MLA) and resulting machine-learning models (MLM) are:

- **MLA:** Decision tree → **MLM:** Logical tree of IF-THEN statements with *specific* values which depend on the training data.
- **MLA:** Linear regression → **MLM:** model of a vector of coefficients with *specific* values which depend on the training data.
- **MLA:** → **Artificial neural networks** (for 'deep learning' i.e. learning within 'deep' or hidden layers of the network, located between input and output layers; a widely used mathematical approach is gradient descent\*) → **MLM: Deep learning models** learn patterns in data by adjusting internal parameters (weights and biases), typically represented as vectors or matrices. These weights are optimized during training using methods such as *gradient descent*, and their values depend on the → training data

\*) Gradient descent is a method for unconstrained mathematical optimisation, proposed in 1847 by Augustin-Louis Cauchy and the most common method to optimise → **artificial neural networks** during 'deep learning' (Ruder, 2017). See → **objective function**.

#### Term relationship:

Related terms:

- Computer program

## Machine learning

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

### Concept description

Machine learning (ML) refers to

- 1) the field of study concerned with ML programs or systems
- 2) A set of computational approaches that train a → machine learning algorithm through the use of → training data in order to create a → machine learning model that is able to generalise to unseen data (→post-deployment input data) for solving a specific real-world problem.

Thus, ML does not require the explicit programming of an algorithm (e.g. by a human actor) but relies on the automated optimisation of internal parameters (→ parameter and hyperparameters) based on data processed by the → machine learning algorithm.

Training machine learning algorithms to retrieve accurate and safe machine learning models is not a trivial task. Problems and pitfalls can be categorised according to four categories:

- **Lack of conceptual relevance:** the basic assumptions and/or the understanding and knowledge of the real-world problem to be tackled by the model are insufficient. Developing a performant model without a robust framing of the problem is likely to fail.
- **Irrelevant data:** The data are not relevant in view of the problem, e.g. not sufficiently representative or simply irrelevant or incorrect.
- **Modelling issues** (see also → algorithm-to-model transition): These include various pitfalls, *inter alia*:
  - *Underfitting*: the model has not been sufficiently trained or is too simplistic or not well designed for the task at hand, and does not perform.
  - *Overfitting*: model has been too tightly trained on the basis of the training data and fails to generalise to new/unseen data. This impedes → generalisability.

### Explanatory note

Typically, during machine learning a → objective function is optimised. A frequent approach for optimisation is gradient descent. There are many objective functions to choose from and choosing the right one in light of the problem to be addressed is an important part of designing an algorithm. Additionally, the selection of appropriate → evaluation metrics is essential to guide model development and assess performance (→ algorithm-to-model transition).

The optimisation aims at addressing or solving a given real-world problem, e.g. discovering relationships or semantic meaning that are hidden in (large) data sets. The resulting model can be used for making predictions from data that it has not been exposed to before, but which draw from the same distribution as the data used for training the model.

### Term relationship:

- Machine learning algorithm
- Machine learning model
- Training data
- ML versus statistics
- Artificial neural networks
- Deep learning
- Algorithm-to-model transition

## ML versus statistics

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

### Concept description

While the science of AI and the realisation of AI technology through → machine learning (ML) draw on mathematics and statistics, AI and statistics are not congruent and statistical approaches should not be labelled as AI and vice versa.

While ML generates statistical inference models, machine learning is not simply a sub-field of statistics. Both statistics and ML can be used for inference (a particular focus of statistics) and prediction, but their approaches differ.

However, only AI employs general-purpose machine → learning algorithms to transform these algorithms, after training/learning into a → machine learning model. These are particularly useful for working with 'wide data', where the number of input variables exceeds the number of subjects.

### Explanatory note

The concept description is based on Bzdok et al (2018) who provide a concise explanation of key differences of machine learning versus statistics. The article by Breiman (2001) touches on two fundamental approaches in statistical sciences, i.e. stochastic data models versus algorithmic models treating the data mechanisms as unknown. See also the comment by Faes et al. (2022).

### Term relationship:

Related terms:

- Machine learning
- Algorithm-to-model transition

## Machine learning algorithm

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

### Concept description

Machine learning algorithms provide a type of automatic programming where → machine learning models represent the resulting programme. The field concerned with developing such algorithms as well as AI systems using such algorithms is referred to as → machine learning (ML) or machine intelligence (Brownlee, 2020; Sarker, 2021a; Sarker, 2022).

More specifically, a → machine learning algorithm is an algorithm that is trained with → data to render a → machine learning model. Machine learning algorithms are algorithms that automate analytical model building by automated adaptation ("learning") of relevant → parameters.

This optimisation is achieved by → objective functions, also called loss functions or cost functions, which provide a measure between predicted and desired outputs. The larger the measured 'gap', the greater obviously the error of the model. Derivatives of these functions are used by an 'optimiser' algorithm (e.g. gradient descent) to adapt the model → parameters during training in an iterative and step-wise process until → model performance is acceptable, e.g. a minimum level of → accuracy has been achieved. In practice, accuracy will need to be balanced against → interpretability (see → objective function).

In summary, the cardinal *objective* of training of the → machine learning algorithm can be understood as minimising the loss and/or functions (i.e. the error of the model) and to maximise hence its capacity to provide useful → output and output data.

Although adaptation happens usually during training the algorithm (e.g. “fine-tuning”) with → training data, it might extent to post-deployment phases using → post-deployment input data in case of → continuous and adaptive learning or ‘lifelong’ learning.

#### Explanatory note

N.A.

#### Term relationship:

Related terms:

- Machine intelligence
- Machine learning
- Machine learning model
- Objective function
- Parameters
- Training data
- Artificial neural networks
- Deep learning

## Machine learning model

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

#### Concept description

A “model” in machine learning is typically understood as the output of a → machine learning algorithm “run” on or “trained” with data and, specifically data → features (attributes) that are conceptually or scientifically relevant for the real-world problem to be modelled (based on [Brownlee, 2020](#)). According to this view, a → machine learning model represents the totality of what was learned by a machine learning algorithm. Any machine learning model is a → algorithm. Use of appropriate data → features that are of ‘high quality’ in relation to the real-world problem to be modelled (→ data quality; → data quality metrics) is a key step in building robust and performant models.

However, developing a model (→ model development) encompasses obviously much more than running a machine learning algorithm on data. It involves a process of making numerous decisions that all influence the quality and performance of a model. We refer to this process as → algorithm-to-model transition. Aspects include for instance choice of the → objective function(s), → model calibration or → evaluation metrics.

Notably, the final machine learning model may be ‘frozen’, i.e. not subject to further machine learning. Alternatively, it may be designed to adapt (i.e. ‘learn’) in response to the processing of → post-deployment input data: see → continuous and adaptive learning.

#### Explanatory note

##### **Models and machine learning models**

Generally, a model is a reduced-scale representation of a (usually real-world) object or system or problem; a model can include a representation of relevant properties of a system, its constituting entities as well as relationships, including of a causal nature.

While AI systems may predict causal relationships, it is not always a given that correct outcomes are based on the right reasons: an AI system may use inadequate features for producing its correct output ([Weld & Bansal, 2019](#)), e.g. → **shortcut learning**. Examining how and why AI models produce an output is thus critical for safeguarding that the modelled ‘logic’ corresponds to the known logic of the real-world problem (→ **interpretability and explainability**). → **Causal machine learning** may represent an avenue for ‘encoding’ the ‘right’ causality in a model from the very beginning.

### **Building a model**

A highly simplified summary of model building encompasses six steps: 1) data collection and processing, 2) → model development, 3) → model validation, → model testing, → model evaluation (including post-processing steps, e.g. precise threshold definition for a classifier, improvements of → model performance, reduction of → bias), 4) deployment and → post-deployment monitoring, or post-market surveillance (see → post-market surveillance, market surveillance, corrective action) in case of → AI-enabled medical device software.

### **Term relationship:**

Related terms:

- Machine intelligence
- Machine learning
- Machine learning algorithm
- Objective function
- Parameters
- Training data
- Artificial neural networks
- Deep learning

## Non-learning algorithm

### **Cluster:** B.8 Algorithm, model, algorithm-to-model transition

### **Concept description**

We refer with non-learning algorithms to ‘classical’ → algorithms that are not designed to automate adaptation of algorithmic → parameters but execute, following a data input of certain modality/modalities, a set of instructions to generate an output (→ output and output data) of a specific modality/modalities.

Such algorithms are explicitly programmed by humans or by artificial intelligence models, generating machine-readable code.

### **Explanatory note**

Non-learning algorithms include knowledge bases that employ a set of rules, e.g. based on IF-THEN logical connections or algorithms using the mathematical approach of fuzzy logic and inference engines ([Sarker, 2022](#)) (→ AI technique, → AI typology). The boundaries between non-adapting and adapting algorithms in practice may be fluid and dynamic. The distinction is nevertheless useful since it defines whether and to which extent training data are required for creating or optimising an algorithm employed in an AI system.

### **Term relationship:**

Synonym:

- Algorithm

Related term:

- Machine learning algorithm
- AI technique
- AI typology

# Objective function

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

## Concept description

### **Objective function**

The process of → machine learning implies the automated adjustment or optimisation of algorithmic → parameters to yield ‘optimal’ desired outcomes, such as a sufficiently close match between *predicted outcomes or values* and *actual values or outcomes*.

The automated optimisation of the → machine learning algorithm yields a ‘model’ (→ machine learning model), which, in simple terms, is an algorithm that is capable of mapping accurately the input data to desired outputs (→ output and output data), such as predictions or decisions. Against this background, the term ‘objective function’ has at least two semantic perspectives:

- *Objective as in aim or purpose:* Simply put, the process of optimisation requires a *mathematical function*. The function that achieves this optimisation aim or *objective* (e.g. minimisation of model error or regularisation to avoid → overfitting (Giordano, 2020) or tackle other constraints) is the ‘objective function’.
- *Objective as in un-biased:* In addition, the objective function provides a purely mathematical measure of model performance at each step of training, providing a measurable value of performance that can then be used for further optimisation. In that sense, the function can be regarded as ‘objective’, i.e. not influenced by human preferences, biases or → heuristics.

### **Loss function and cost function**

Main examples of objective functions in machine learning are the ‘loss function’ and the ‘cost function’:

- The *loss function* evaluates individual predictions: it calculates the distance between a *predicted output* (i.e. single prediction) and the desired output for a given example (see → feature) of the → training data. One could therefore call the loss function also ‘single example error function’.
- The *cost function* typically aggregates the loss over subsets of the → training set (e.g. by calculating the *mean square error* across examples). The number of subsets is called ‘batch size’. Averaging loss over the *entire* → training data set is typically not feasible with large data sets. The cost function reflects the overall error of the model across (subsets of) examples (see → feature). One could call the cost function also ‘dataset error function’.

The objective function, over iterative learning runs (see → hyperparameters) guides the adjustment of the model’s intrinsic → parameters, thus leading to an improved match between the model’s predicted values and the actual values.

### **Optimiser**

Notably, adjustments of → parameters to minimise typically the ‘cost function’ are not made by the objective function itself, but by the so-called ‘optimiser’, using derivatives of the objective function. Optimisers vary in regard to how they adjust parameters. Thus, the choice of optimiser will influence the model optimisation process and ultimately model performance. Optimisers can be loosely categorised in two groups (Giordano, 2020): fixed gradient descent (GD) and adaptive optimisers – most of them are based on gradient descent. Particle swarm optimisation is an example of a non-GD optimiser which is less frequently used in machine learning though.

- **Fixed gradient descent (GD)** is a well-established algorithm for unconstrained mathematical optimisation (see e.g. Wikipedia, 2024). Updates are made in the direction of the steepest descent (see e.g. IBM, 2024 for an introduction to gradient descent in machine learning). Use of gradient descent-based optimisers involves the manual tuning of → hyperparameters such as step size or learning rate. This can be considered a drawback. An automation of GD learning

rate tuning has been proposed by [Chandra et al., 2019](#). Concrete approaches include batch GD, mini-batch GD and stochastic GD (SGD).

- **Adaptive optimisers** automate the tuning of learning rate.
  - ‘**RMSprop**’ (=root mean square propagation) belongs to a group of methods using adaptations of learning rate (→ **hyperparameters**), originally proposed by Hinton and co-workers ([Hinton et al., 2018](#)). RMSprop It is an extension of Stochastic Gradient Descent (SGD) and the ‘momentum method’; it is the basis of the ‘Adam’ optimiser algorithm ([Huang, 2020](#)). For an introduction to RMSprop see [Kashyap, 2024](#); for a technique to improve RMSprop efficiency ([Elshamy et al., 2023](#)).
  - ‘**Adam**’ involves the adaptation of learning rates (→ **hyperparameters**) for each → **parameter**, utilising preceding gradients.
  - Other examples are **Adagrad** and **Adadelta** ([Giordano, 2020](#)).

### **Examples of loss functions**

#### Binary classification

- ‘cross entropy loss’ is the most commonly used loss function for medical image analysis tasks, e.g. diagnostic prediction / disease detection; [Rajamaran et al., 2021](#).
- ‘hinge loss’ is mainly used in support vector machines,
- ‘focal loss’ can address detection challenges, e.g. detecting small lesions in a medical image.

#### Multi-class classification

- Categorical cross entropy (CCE)
- Sparse categorical cross entropy (SCCE)
- PolyLoss

#### Regression

- mean square error (MSE),
- mean absolute error (MAE) or
- combinations of the two (also called ‘Huber loss’).

#### Probabilistic models

- maximum likelihood estimation
- negative Log-likelihood

#### Next-token prediction loss

- Next-token prediction loss (NTPL). In LLMs a ‘token’ is typically a word, a syllable or any other part of a word. NTPL measures how well the LLM predicts the next ‘token’ in a sequence of previous tokens, i.e. ‘generates’ the next likely word when composing a text based on contextual information provided in the prompt.

Computer vision (image segmentation, registration, object detection)  
MSE. Cross-correlation, cross-entropy, Dice loss

N.B. generative visual AI models that predict the next pixel based on contextual information use other objective functions that can be grouped as per-pixel loss functions. These include MSE, MAE, CCE. Per-pixel loss functions are relevant also for other tasks of visual AI / computer vision, including object detection and image segmentation.

### **Choice of objective function: balancing model accuracy and model interpretability**

The objective function has a considerable impact on both → **accuracy** and → **interpretability** and **explainability** of a model. For instance, functions differ in respect to → **model performance aspects**. As an example, MSE ‘penalises’ large errors more strictly than the MAE. This influences accuracy. There

may be differences concerning the robustness of algorithms concerning noise or outlier values in the data set. Generally, the simpler the objective function (loss or cost function), the more intrinsically interpretable the model (e.g. MSE of linear regression). More complex functions optimising various terms may render more accurate models at the cost of interpretability, since optimisation of composite objectives makes it more difficult to determine which aspect influenced a particular decision. Thus, the inherent trade-off between accuracy and interpretability hinges to a large extent on the choice of objective function. Another important aspect to consider is regularisation (see → overfitting).

### Explanatory note

For an introduction to loss functions and optimisers for deep learning, see [Miller, 2022](#).

For a review of loss function in deep learning see [Terven et al., 2023](#).

### Term relationship:

Related term:

- Machine learning algorithm
- AI technique
- AI typology

## Parameters

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

### Concept description

Parameters are model variables that are adjusted during machine learning, while → hyperparameters are settings or configurations chosen by the person supervising and designing the machine learning process.

→ Machine learning models are ‘parameterized’, i.e. they feature ‘variables’ or ‘parameters’ that are adjusted during → machine learning, when the → machine learning algorithm is run on → training data. Examples include:

- **weights** that determine the strength of a given input to a virtual neuron in an → artificial neural network.
- → **neural network bias**, influencing the positioning of the activation function along the input axis of a virtual neuron within an artificial neural network.
- **coefficients** in linear regression, such as slope and intercept.
- **decision boundaries** in decision-trees.

### Explanatory note

N.A.

### Term relationship:

Related terms:

- Neural network bias
- Objective function
- Machine learning
- Machine learning algorithm
- Machine learning model

## Hyperparameters

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

### Concept description

Unlike → parameters, hyperparameters are **not model variables** but training settings chosen by the machine learning supervisor (e.g. computer scientist, data scientist) before starting the → machine learning process. Examples:

- **learning rate** – a key setting which sets the pace at which the → machine learning algorithm adjusts parameter estimates (→hyperparameters) and/or sets the precise values for these parameters (e.g. weights, → neural network bias etc.). Learning rates are not kept constant but subject to ‘learning rate decay’  $\alpha$ : The learning rate typically is higher at the beginning of training in order to accelerate finding a near-optimal solution (see → optimisation function, e.g. gradient descent), but is then decaying in order to avoid unstable or oscillation-like behaviour. Setting the right decay rate and thus finding the sweet spot between initial speed and appropriate slowdown is crucial for good results in deep learning. There are a range of other approaches, e.g. adaptive learning or cyclic learning.
- **number of hidden layers** in a → artificial neural network. Generally, a higher number of hidden layers will afford the model greater capacities to capture more complex patterns.
- **number of training runs**, steps per epoch.
- **strength of regularisation** lambda (see → overfitting).
- **dropout rate** (see → overfitting) which determines the fraction of neurons that randomly ‘drop out’ at each iteration of training a → artificial neural network.

### Explanatory note

Learning rate: for an accessible introduction, see [Mishra, 2023](#). For a detailed overview see [Gonsalves & Upadhyay, 2021](#).

To lower the hurdles for non-expert users, **automated hyperparameter optimisation (HPO)** is increasingly used ([Yu & Zhu, 2020](#)).

### Term relationship:

Related terms:

- Parameters
- Neural network bias
- Objective function
- Machine learning
- Machine learning algorithm
- Machine learning model

## Neural network bias

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

### Concept description

Neural network bias (NNB) is an important variable or → parameter of → artificial neural networks. NNB influences at what level of summated inputs an artificial neuron activates and produces an output – which in turn acts as an input to other connected neurons in the layer. The bias influences the output of the activation function, enabling the network to better fit the data.

NNB acts by shifting the activation function of an artificial neuron (e.g. a sigmoid function; see → deep learning) along the input axis (e.g. x-axis) by adding a constant (the specific ‘bias’) to the input.

NBB is typically added to the weighed input of a neuron before the activation function is executed ('weights' are another important → parameter in → artificial neural networks; see → deep learning).

Another key influence on weights is the learning rate, which is not a model variable but a → hyperparameter.

#### Explanatory note

For introductions to neural network bias and the activation function, see [Zouari, 2021](#) and [Turing website, 2025](#).

#### Term relationship:

Related terms:

- Artificial neural network
- Deep learning
- Parameters

## Artificial neural networks

### **Cluster:** B.8 Algorithm, model, algorithm-to-model transition

#### Concept description

Artificial neural networks (ANN) are an abstract and highly simplistic computational representation of biological neuronal ensembles (D. Hebb's 'cell assemblies'; [Hebb, 1949](#)) and – at the present stage of ANN development – of localised cortical networks or circuitries. While ANNs do not recapitulate such assemblies, their basic design is rooted in these at an extremely reduced level of complexity.

For example, the ANN (especially CNNs or vision transformers) feature of an ‘input layer’ that projects to several “hidden” or “deep” layers can be seen as rooted in knowledge about localised cortical circuitry: visual input from the retina and visual thalamus project to layer IV granular neurons that represent the primary “input layer” of visual cortex. These neurons in turn project and converge on pyramidal cells in layers II and III. These cells have extensive synaptic connections with each other within these layers and project to other secondary visual cortical areas ([Braitenberg & Schüz, 1991](#)). The adjustability of the weight of input of ANN neurons can be seen as rooted in the high extent of synaptic plasticity (a substrate for learning, e.g. of network properties, including sensory map development, sensory feature extraction) within cortical layers.

Despite the extreme simplicity of their highly modular setup (repetitive arrangement of artificial neurons that are connected to other neurons in proximal ‘layers’; adjustment of connective strength via ‘weights’), ANNs are very powerful tools for a wide range of tasks, e.g. detecting and learning of features in data, such as orientation of edges, facial features, anatomical features etc. and are extensively used for medical image analysis or predicting the probability of the next word (in a large language model) or the appropriate next pixel or image properties (in a → foundation models optimised for image synthetisation; see → generative AI).

#### **A very brief history of ANNs**

One of the earliest and simplest ANN was Frank Rosenblatt's ‘perceptron’ in 1958 ([Rosenblatt, 1958](#); [Block, 1962](#); [Lefkowitz, 2019](#)), a simple one-layered neural network with adjustable network biases (→ neural network bias) regulated by potentiometers. Limitations of this architecture were pointed out by [Minsky & Papert, 1969](#) (reviewed by [Block, 1970](#)), leading to the first “AI winter”, i.e. a decrease of enthusiasm and interest in ‘learning machines’.

With the advent of '**computational neuroscience**' (see the 'programmatic' paper by Sejnowski et al., 1988, following the seminal Carmel conference in 1985), computational models of neural networks have been and continue to be used in neurosciences as a research tool (Hopfield, 1982) - for instance to test hypotheses of receptive field or sensory map formation, e.g. in primary visual cortex (e.g. Heeger et al., 1996; Carreira-Perpinan et al., 2005) For an introduction to computational neuroscience, see Wang et al., 2020.

However, ANNs came to a wider attention only in the 1980ies with the arrival of large-scale and highly performant convolutional neuron networks, CNNs (a form of ANNs), e.g. ConvNet (1989), LeNet (1998), AlexNet (2012), ResNet (2015), up to EfficientNet (2019/2020) (see Wang et al., 2023). A major push for ANNs as an AI technique for addressing 'real-world' problems' came with AlexNet (a CNN) winning the ImageNet visual recognition challenge in 2012, a success that was, according to the authors of AlexNet, due to the number of layers (→ **deep learning**) of the model (Krizhevsky et al., 2012).

### **Artificial neural networks, McCulloch-Pitts neurons and synaptic plasticity**

Each layer in a neural network is composed of artificial 'neurons' that can be seen as a basic and highly simplistic recapitulation of biological neurons (spearheaded by hypothesized concept such as Labique's 'integration and fire' neuron and McCulloch – Pitts' neurons (see → **computational theory of mind**; → **AI technique**). Such neurons are based on neurobiological findings and theories of information integration in neurons and synaptic plasticity, i.e. the malleability of neuronal connection strength based on neuronal (co)activation, first postulated by the psychologist Donald Hebb in 1949 ("cells that fire together, wire together". This malleability, confirmed in countless neurobiological, -physiological and -anatomical studies is postulated as the plausible substrate for feature map formation in cell ensembles as well as memory formation within the cortico-hippocampal brain circuitry or the formation of 'mirror neurons' with predictive properties (Keyers & Gazzola, 2014).

### **Artificial neural networks as an abstraction of biological neurons in cell ensembles (e.g. visual cortex)**

The basic outline of ANNs have been heavily influenced by research in cortical sensory development and especially that of primary visual cortex (brain area 17). Cortical neurons and cell ensembles in primary visual cortex are involved in visual feature extraction such as ocular dominance or orientation selectivity, basic 'skills' for analysing visual scenes and recognising meaningful features (e.g. object, faces). There are various plasticity processes that drive the development of visual feature maps (see for instance Li et al., 2023; Cline et al., 2023; Sun et al., 2019; Tanaka et al., 2020; Müller & Griesinger, 1998): synaptic long-term potentiation and structural plasticity, i.e. the anatomical strengthening as well as physical 'pruning' of connections between neurons (for an elegant theory of map formation, see Najafian et al., 2022). Each 'functional column' in the primary visual cortex is a neural network - with layer IV receiving visual inputs (from retina and thalamus), feeding these forward to middle layers II and III which 'process' this information via extensive connections within these layers, and 'outputs' to other cortical areas being triggered in layer VI. Artificial neural networks are a highly simplified abstraction of such primary sensory cortical architecture.

### **Artificial neurons: inputs, weights, summation function, activation function and bias**

- Artificial neurons receive **inputs** from other artificial neurons and connect to other neurons in the various layers of the network.

*This can be seen as a highly simplistic model of synaptic input on real neuronal dendrites and cell bodies, both excitatory (membrane depolarising) and inhibitory (membrane repolarising).*

- These inputs can be **weighed** (by values called 'weights').

*This is a simplistic model of processes of biological synaptic potentiation and structural plasticity of synaptic connections (see for instance Li et al., 2023; Cline et al., 2023; Sun et al., 2019; Tanaka et al., 2020; Müller & Griesinger, 1998).*

- A so-called '**summation function**' integrates these weighed inputs.

*This mimics dendritic and somatic integration of excitatory synaptic input in biological neurons. The → **neural network bias** shifts the activation function along the x-axis, influencing from which summated input level a neuron activates.*

- Finally, the so-called '**activation function**' will determine at which level of excitation an artificial neuron will produce an output signal. Activation functions may be non-linear and 'transduce' the linear regression of input signals into an output.  
*This function is an abstract representation of the biological factors influencing the production of an action potential at a neuron's axon hillock in response to summated depolarisations caused by net synaptic excitation received at dendritic and somatic synapses. The action potential likelihood is determined inter alia by the density and activation state (e.g. phosphorylation state) of voltage-dependent membrane channels at the hillock.*

An important aspect of deep learning is the shifting of the activation curve through so-called → **neural network bias**. Together with the shape of the activation function curve, bias recapitulates the readiness of a neuron to fire an action potential as determined by both state and density of voltage-sensing ion channels at the axon hillock.

#### Explanatory note

N.A.

#### Term relationship:

Related terms:

- Artificial neural networks – types
- Deep learning
- AI technique
- Artificial intelligence (AI)

## Artificial neural networks – types

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

#### Concept description

There are various types of → **artificial neural networks** with applications in health and medicine, including

**Multi-Layer Perceptrons (MLPs)** can be used for several classification tasks (e.g. [Bikku, 2020](#); see also [Rosenblatt, 1958](#) for the original "Perceptron").

**Convolutional Neural Networks (CNNs)** are particularly suited for image recognition and processing. They are used widely in medical imaging, e.g. for diagnostic and prognostic predictions, disease detection etc. based on radiological images (for a review: [Wang et al., 2024](#)).

**Recurrent Neural Networks (RNNs)** and **Long Short-Term Memory (LSTM)** are useful for processing sequential data and time series data, e.g. for disease progression (prognostic predictions) or monitoring of clinical parameters/vital signs. They may be valuable for analyzing electronic health records (EHRs) (see for instance: [Rasmy et al., 2021](#); [Olaimag & Bozdag, 2024](#)).

**Autoencoders:** these are particularly suited for dimensionality reduction and feature learning for identifying patterns in large medical datasets, for genomics research (gene expression patterns; see [Khan et al., 2025](#)) and for anomaly detection in medical data and identifying rare conditions. For example, U-Net is a CNN with an autoencoder structure that is commonly used for medical image analysis tasks.

**Deep Belief Networks (DBNs)** are suited for learning hierarchical representations and have been used for disease diagnosis, disease characterization (e.g. [Pinaya et al., 2016](#)) and disease risk prediction as well as drug-drug interactions.

**Generative Adversarial Networks (GANs)** – used for generative tasks, e.g. the creation of synthetic medical images for training purposes and to augment datasets, e.g. for rare diseases (→ [synthetic data](#)) ([e.g. Little et al., 2021; Arora & Arora, 2022](#)) or for image classification (e.g. [Wang et al., 2021](#)).

**Transformer Networks** are suited for processing sequential data with ‘attention’ mechanisms (‘attention weights’; see → [deep learning](#); → [foundation models](#)). Such networks can be used for various tasks, e.g. medical information processing, transcribing/analyzing clinical including surgical notes ([Nerella et al., 2024](#)) and support health research (e.g. protein sequences and structural biology predictions, medical image analysis ([Zhu & Wang, 2023](#)), image segmentation ([Zhang et al., 2024](#))). Transformers or transformer networks can refer to both: an element of AI architecture as well as a model type in case of complete trained systems).

#### Term relationship:

Related terms:

- Artificial neural networks
- Deep learning
- AI technique

## Deep learning

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

#### Concept description

‘Deep learning’ refers to → [machine learning](#) using → [artificial neural networks](#) with hidden layers. The term was originally introduced by [Dechter \(1986\)](#).

Artificial neural networks are typically composed of three types of layers, with the number of layers of artificial neurons varying: 1) input layer, 2) hidden layer(s), 3) output layer.

The term ‘deep learning’ is an allusion to the ‘depth’ of the hidden layers in → [artificial neural networks](#) ([Sarker, 2021b](#)). Note should be taken however that ‘learning’ (i.e. adaptation of algorithmic parameters) in such networks is not necessarily restricted to the hidden layers, but affects weights and biases between every layers. Learning affects model parameters (→ [parameters](#)) and is critically shaped by the setting of → [hyperparameters](#) (e.g. learning rate, number of training epochs or runs).

Parameters in → [artificial neural networks](#) include *weights*, i.e. the strength of connections from neurons in layers proximal to the neuron of interest, *bias* (→ [neural network bias](#)), a constant added to the sum of all weights (summation function) prior to execution of the activation function; *convolutional kernels*: learnable filters that can identify features or patterns applied to processed input data in view of feature detection; *attention weights*: used in transformer-based artificial neural network architectures (→ [foundation models \(FM\)](#); → [generative AI](#)) that support learning of specific parts of an input (e.g. query) should be weighed (hence ‘attention’) ([Vaswani et al., 2017](#)).

#### Further reading

For an introduction, see [Faisal, 2020](#).

For reviews of dep learning for prediction models: [Emmert-Streib et al., 2020](#).

[Talaei Khoei et al., 2023](#) provide a systematic review of deep learning challenges and research directions.

#### Term relationship:

Related terms:

- Artificial neural networks
- AI technique

- Foundation models (FM)
- Generative AI

## Algorithm-to-model transition (ATMT)

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

### Concept description

We propose a conceptual framework for considering all key decisions to be made when transiting from algorithm and training data to a model that relates (→ conceptual relevance) to a real-world aspect of our world. We call this process ‘algorithm-to-model transition’ (ATMT).

ATMT encompasses the totality of *assumptions* and *decisions* made during the trajectory from a → machine-learning algorithm run on → training data (including pre-trained algorithms) to a → machine-learning model that is tailored to a specific real-world problem or has broader applicability to various problems (→ foundation model, → generative AI).

Relevant assumptions and decisions during ATMT should be systematically collected and documented for reasons of → traceability, for internal and external → failure transparency (→ see TRANSPARENCY) and to allow interrogating how specific decisions may have caused problems – a prerequisite for evidence-based improvement. ATMT and failure transparency also impact on the ability to appropriately respond to errors or problems (see → responsiveness, contestability, redress, liability).

ATMT may employ classical → non-learning algorithms but will typically involve a → machine learning algorithm to create a → machine learning model. Importantly, ATMT is an entry point for possible → biases that can cause → algorithmic bias. Decisions taken during ATMT impact on the → intelligibility of the AI system and its outputs (e.g. through choices including → AI technique and complexity of a given model, e.g. pretrained convolutional neural network).

- **ATMT relates primarily to three life cycle states** (see → life cycle of AI in health) of
  - concept, planning and design
  - development and, to a certain extent,
  - validation.
- **Life cycle processes:** ATMT builds primarily on life cycle processes, but should *not be equalled* with procedural aspects. ATMT is about **building awareness** of all steps in a **forward-looking way during design stages as well as during the subsequent trajectory**. This involves collecting necessary evidence on data collection/wrangling and modelling decisions and outlining how these decisions were motivated by scientific and/or clinical considerations.
- **Value chain elements:** ATMT may touch on value chain elements (→ value chain of AI), i.e. **data** (including from data providers), **models** (included pre-trained models), **enabling IT infrastructure, enabling technologies** (e.g. platforms, frameworks, machine-learning libraries) and **cybersecurity** (see [Reina & Griesinger, 2024b](#)), which we consider a value-preserving enabling for other elements of the value chain.
- The general **ATMT concept may be applied in a modular fashion:** ATMT elements that impact predominantly on intelligibility can be dealt with in an “*intelligibility pathway*”, or those that impact on bias in an “*bias mitigation pathway*” in which necessary steps and decisions are planned and collected for internal and, where necessary, external documentation (→ TRANSPARENCY – communities of transparency).

- **Validation and evaluation involve fine-tuning:** Since → model validation and → model evaluation) may typically involve fine-tuning or post-processing of models, these exercises are considered relevant for ATMT.

Outside the scope of ATMT are all those verification, validation and evaluation procedures that do *not* lead to adaptations of the → machine learning model, but focus on specific aspects of the model or AI system. These include usability (→ usability validation), technical validation of the AI system (→ AI system validation) or → clinical validation.

#### **Explanatory note**

Algorithm-to-model transition involves the following aspects (see also the short overview by [Naeem, 2024](#); for a detailed description of AI model development and validation, see [Liu et al., 2022](#)). These need not necessarily be tackled in the sequence below:

#### **Assumptions and evidence for conceptual/scientific relevance and context relevance:**

- A fundamental consideration for building a model concerns the underpinning → conceptual relevance. This related to the validity of relevant (e.g. scientific) evidence, the strength and plausibility of that evidence as well as additional assumptions that may be plausible but not sufficiently backed up yet by available evidence at the stage of conceiving or designing an AI system. For AI systems used for clinical/medical applications, see the concept of → valid clinical association / scientific validity.
- A separate consideration concerns the → contextual relevance, which concerns relevance considerations in regard to the intended → use context and actual use (→ use of AI systems in health and healthcare) as well as the envisaged → use environment (e.g. hospital or home care).
- Incorrect, outdated assumptions may introduce hidden → bias: thus, consideration should be given to *intrinsic bias* in medicine ([Straw, 2020](#)), e.g. traditional assumptions that are not sufficiently questioned or not treated as such but rather as established knowledge, although robust evidence may be missing (see section on → FAIRNESS).

#### **Data**

- **Data collection and data sampling decisions:** Collecting the right data of sufficient quantity and quality to enable training a model and yielding adequate performance characteristics (e.g. accuracy, precision), to validate the model and to assess or evaluate it. Aspects of → data quality and → data quality metrics need to be considered. Choosing → proxies and → attributes in view of the problem to be tackled. Data collection decisions refers also to situations where data are purchased by data providers of the → value chain of AI. An important source of bias at this stage stems from **data sampling** (→ sample / sampling): if the patient group or cohort of which data were sampled for training a → machine-learning algorithm is not representative of the population in which the final → machine-learning model will be used, this sample bias will be propagated in real-world use in a healthcare setting ([Adamson & Smith, 2018](#)). Thus, demographic and baseline characteristics (e.g. gender, race, ethnicity, number of patients, age) should be considered in view of the → intended use, intended → use environment and → use context (see also → use of AI system in health and healthcare). These aspects should be transparent so that end users can gauge potential constraints of an AI system. Sample bias can be considered a robustness issue (→ risks related to insufficient robustness / resilience). Reporting checklists such as PROBAST-AI (see Annex I) should be consulted.
- **Data processing decisions:** preparing data sets e.g. by removing duplicate data, identifying outliers, closing data gaps (→ data processing / wrangling). **Correctly labelled data** (where required) are crucial. Incorrect data labelling can introduce → bias and, subsequently, performance and safety issues.

## **Algorithm / model**

- **Model selection decisions:** e.g. selection of an → AI technique (e.g. neural network, decision tree, random forest) and → machine learning algorithm that is suitable for the problem to be tackled by the final AI system and its intended → AI system output and tasks. Model selection also entails the use of ready-made models and pre-trained models that are subsequently trained by → transfer learning. Selecting a pretrained model (e.g. ResNet) and selecting the required level of complexity (e.g. number of layers).
- **Decisions concerning the balancing of performance versus interpretability / explainability and model complexity:** some models may allow intrinsic interpretability, potentially at the cost of performance ([Kaddour et al., 2022](#); see however [Burkart & Huber, 2021](#)), while other models (e.g. neural networks) will require more onerous post-hoc interpretability or explainability techniques (→ interpretability and explainability) to understand how and why outcomes have been reached. Thus, the → intended use, → use context and also → use environment will have to be considered when choosing a model: is performance the top priority and is thus more effort to obtain interpretability justifiable or is intrinsic interpretability more important so that a relative drop in (predictive) performance may be acceptable? Given the need for **informed consent by patients** (→ ensuring the means for free and informed consent), → traceability and understandability of errors (including systematic ones, i.e. → bias), interpretable models should always be considered first ([Rudin, 2019](#)). Only in case such models cannot address the real-world problem satisfactorily, ‘black box’ models (e.g. requiring post hoc explainability) should be considered (see comments by [Rudin, 2019](#); [Vokinger, 2021](#)). Consideration should also be given that for specific applications (especially decision-making) “simpler”, intrinsically interpretable models may be just as performant as “block-box” AI types which require more effort in regard to explainability (see for instance [Soliman et al., 2023](#)). Finally, performance requirements (e.g. in view of data to be analysed) versus model complexity needs to be balanced (it may be preferable to choose comparably simpler neural networks over more complex ones for a specific task (e.g. image analysis/computer vision).
- **Decisions during training and modelling, evaluation metrics including model calibration (where applicable):** This includes running the → machine learning algorithm on the processed data, setting → hyperparameters (e.g. number of training epochs; see → [hyperparameters](#)), setting thresholds; choosing → evaluation metrics (e.g. recall, precision) for → model performance assessment including decisions on → model calibration (where applicable). A key decision concerns the choice of the appropriate → objective function (or loss function) or other constraints to be considered when optimising the model ([Terven et al., 2023](#)).
- **Decisions concerning techniques for fairness, non-discrimination, explainability, de-biasing and bias mitigation:** many approaches for explainable AI (XAI) are available under the umbrella of the FAT ML and XAI communities ([Veale & Binns, 2017](#); [Selbst et al., 2018](#)). Google, Microsoft and Facebook have issued general toolkits, e.g. AI fairness 360”, “What-if-tool”, “Facets”, “fairlern.py” or “FairnessFlow” ([Whittaker, 2018](#)). Explainability in healthcare has been examined in detail by various authors. Holzinger and colleagues have proposed the new useful concept of ‘→ causability’ to address the quality of explanations for medical purposes (e.g. [Holzinger et al., 2019, 2020](#); [Müller H et al., 2022](#) in the context of image analysis and in vitro diagnostics). Bias mitigation methods are available ([Tejani et al., 2024a](#)), including mathematical de-biasing approaches such as adversarial de-biasing ([Zhang et al., 2018](#)) or oversampling ([Kamiran & Calders, 2009, 2012](#)) have been proposed. In addition, → continuous and adaptive learning has been proposed as an avenue to reduce bias during post-deployment, as long as rules acquired during the training phase are not lost ([Lee & Lee, 2020](#)).

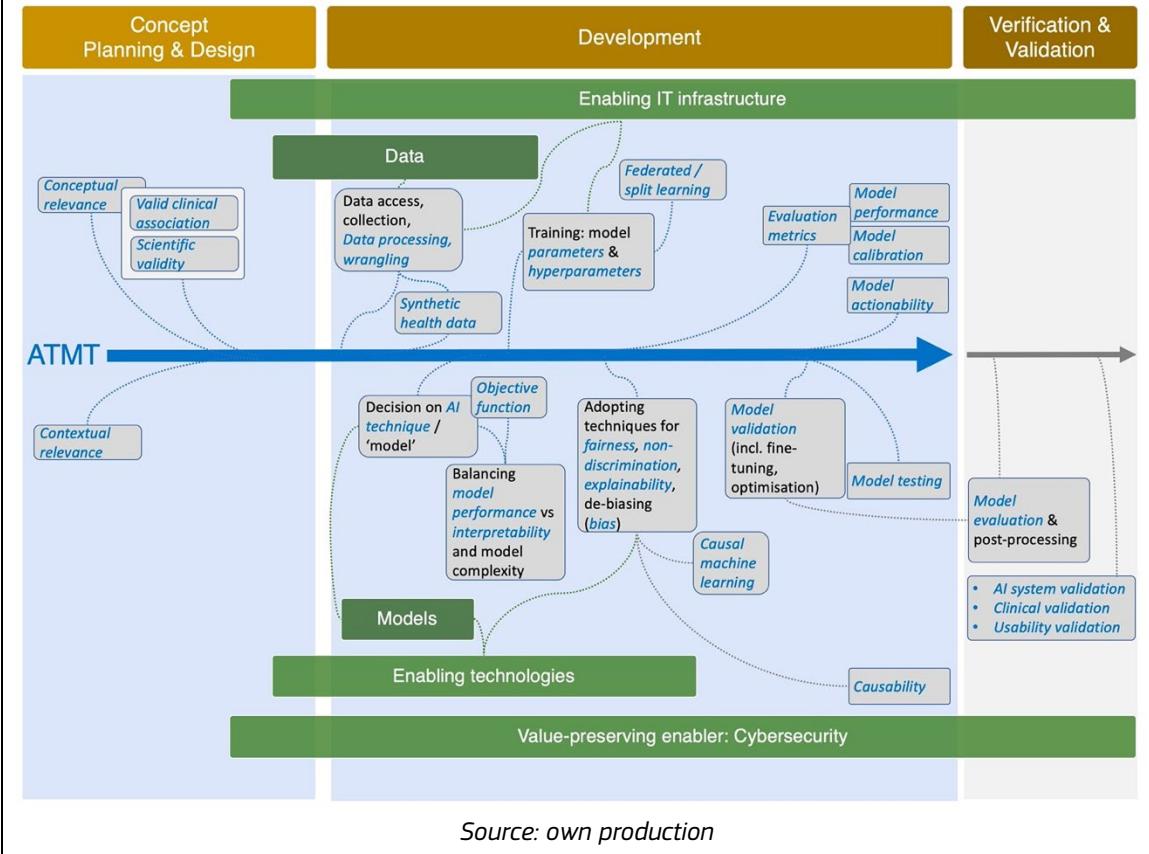
- **Decisions during fine-tuning, optimisation, improvement**, performed during → model validation, → model testing or → model evaluation. This may involve threshold values for predictors, de-biasing.
- **Decisions concerning ‘causability’ and ‘model actionability’**: → causability, i.e. the usefulness and adequacy of explanations for medical purposes and ‘→ model actionability’, i.e. considerations of the value of a model in a clinical decision-making context, especially when compared to a human decision maker (see also → BENEFICENCE – gains).

#### Term relationship:

Related terms:

- Model development
- Machine-learning algorithm
- Machine-learning model
- Model validation
- Model testing
- Model evaluation
- Data
- Development data
- Causability
- Model actionability

**Figure 25.** Schematic depiction of the algorithm-to-model transition pathway (blue arrow) for → machine-learning based models/AI systems. Relationships to life cycle stages (yellow) and value chain elements (green) are indicated (\*see Reina & Griesinger, 2024b). ATMT should provide evidence on assumptions and decisions that are crucial for arriving at the final model. This includes model post-processing or optimisation steps based on validation and evaluation exercises. Other verification and validation steps (e.g. verification of design specifications of AI system, AI system validation, usability validation, clinical validation) are not part of ATMT. Terms in blue italic font correspond to ontology entries.



## Federated learning & split learning

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

### Concept description

Training AI models requires high-quality data. Further, depending on → AI technique (e.g. deep learning) and → AI typology (e.g. multi-modal AI) training a → machine learning model may require large amounts of data.

Traditionally, → training data have been collected by a developer locally, but the necessary data transfers or data transmissions lead to obvious concerns regarding the preservation of data privacy (→ PRIVACY PROTECTION), data ownership. Furthermore, depending on jurisdictions, such transfer may be subject to legal restrictions. Collecting a sufficient amount of health data (that should be free of → bias) is an enormous challenge due to issues of data availability, cost and obvious concerns regarding personal information (→ risk related to data privacy of personal information) protected, for good reasons, by relevant legal provisions. However, this situation should not hinder the development of promising AI models for health applications.

Federated learning and split learning represent potential solutions. Both approaches are based on the ‘model-to-data scenario’ where developers train and test → machine learning models without sharing raw data. These approaches are particularly relevant for health applications that built on personal and sensitive information from people (Dhade & Shirke, 2023; Narmadha & Varalakshmi, 2022; Zhang F et al., 2023).

- Federated learning can be considered an approach tackling “privacy by design”: large → machine learning models are trained across multiple data centres that hold sensitive data, without the need for transmitting any sensitive information. In the area of health, this allows to balance → medical privacy / health privacy with the desirable possibility of tapping on large distributed data sets, e.g. from electronic health records / electronic medical records that may contain valuable information on diseases pathways, risk factors or other medically relevant information. The large model obtained can be subsequently adapted to other (and often more specific) tasks by → transfer learning. Federated learning is sometimes considered a technique of → transfer learning (e.g. Chato & Regentova, 2023).
- It has been argued that ‘split learning’ provides even better privacy (Thapa et al., 2020) by splitting the machine learning model (e.g. an artificial neural network) into segments that run on different servers, e.g. with a first layer operating on a developer’s server, producing so-called ‘activations’ (intermediate). These are sent to another server which completes training without access to or need for the raw data. Federated learning in contrast keeps copies of the full model on each federated server, allowing each participating party to train the model with their local data set.

#### Explanatory note

Federated learning is not a perfect solution in regard to → PRIVACY PROTECTION. For instance, risks of **data leakages** have been reported for federated learning. **Cyber attacks** of malicious actors have led to extraction of private and sensitive training data, e.g. by analysing the machine learning models trained on federated data or by tapping on information exchanged during the federated learning process (Baracaldo & Xu, 2022).

There are also issues in regard to → model performance, → FAIRNESS and → bias resulting from inadequately trained federated models. This can have severe consequences due to their critical nature in the value chain (Sinosoglou et al., 2023).

#### Term relationship:

Related terms:

- Machine learning algorithm
- Training data
- Machine learning (ML)
- Artificial neural networks
- Deep learning

## Continuous and adaptive learning

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

#### Concept description

The concept of using → machine learning algorithms also during the post-deployment stage for the continuous adaptation of the → machine learning model and its → model performance. Continuous and adaptive learning are useful in changing environments. For introductions to continuous learning, see Ferreiro, 2023; Algolia, 2023; Wang et al., 2023; Leena AI labs, 2024.

### **Adapting to drifts / shifts**

Continuous and adaptive learning promises to allow addressing performance challenges resulting from a variety of changed conditions during the post-deployment stage, including → **concept drift**, → **data drift** (covariate shift) or → **distributional drift / shift** (also known as 'deployment bias': see → **bias**).

### **Enhancing robustness and generalisability**

In the context of addressing such drifts and changing conditions during use, continuous learning is also referred to as '*adaptability*'. A second goal of continuous learning is to improve robustness (→ risks related to insufficient robustness / resilience) and → **generalisability** of AI models.

### **Explanatory note**

The concept, as so many in the practice of artificial intelligence, is fluid and still emerging. However, it is clear that continuous learning is intended to overcome the challenges of static or frozen datasets that are employed during the development stage (→ **model development**). Continuous learning can be categorised into two fundamentally different approaches:

- 1) So-called **incremental learning** refers to the process of keeping a model up-to-date with new data, without having to completely re-train a deployed model. This can be considered part of "model maintenance". In healthcare, higher resolution radiological images may require updating a deployed model with these new data. Importantly, equal care regarding the quality of → **training data** as during the development stage needs to be employed. Caution must be taken to avoid that incremental learning leads to undesired → **drift / shift in machine learning**, e.g. in case an AI system is used for analysis of both images of newer and older quality characteristics. Radiologists may need to reassess images of a patient over several years (e.g. slowly progressing neuroendocrine tumours; see [Zheng et al., 2023](#)) and → **drift / shift in machine learning** could severely impact performance.
- 2) In contrast, **lifelong learning**, refers to the concept to train a model throughout its life cycle, including on new and other data. Lifelong learning is a key characteristic of → **foundation models**. Lifelong learning may lead to performance drifts (see: [Chen et al. 2023](#)). Thus, lifelong learning may not be an adequate approach for many AI systems used in healthcare which depend on stability, → **reliability** and repeatability.

### **Term relationship:**

Synonyms:

- Continual learning

## Causal learning

### **Cluster: B.8 Algorithm, model, algorithm-to-model transition**

### **Concept description**

"Causal learning" is a topic of philosophy, cognitive science, psychology, and particularly the sciences of learning. Causal learning is concerned with the fundamental problem of how humans learn about the causal structure of reality. Causal learning can thus be considered a topic of epistemology, i.e. how humans can acquire knowledge and what are the conditions and operations of human understanding.

Briefly, causal learning aims at formulating, developing and testing (probabilistic) model theories of learning and development ([Gopnik, 2024](#)). Over the last quarter century, theories of how the (developing) human mind learns concepts and rules relating to causality have been developed by philosophers of science and computer scientists (e.g. a formal account of causal knowledge and learning, e.g.

[Glymour, 2001](#); [Spirtes et al., 2001](#)) and developmental psychologists (e.g. the “theory theory” emphasising that cognitive development in humans resembles theory formation in science; [Gopnik 2012](#)) or, perhaps more appropriately, *vice versa*.

#### Explanatory note

Causal learning makes use of various approaches to describe causal connections between variables, notably “Bayes nets”, graphical depictions of variables. Bayes nets are part of a wider approach in cognitive science, namely Bayesian probabilistic models of cognition.

#### Term relationship:

Related terms:

- Causal machine learning

## Causal machine learning

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

#### Concept description

Causal machine learning is a term used to describe machine learning methods and approaches that utilise a model of causally connected variables (a so-called “structured causal model”, SCM). While → **machine learning** techniques usually are based on associations between variables in data, causal models and their application to machine learning (causal learning) allow to differentiate between actual causal relations and simple (spurious) correlations. Causal machine learning draws on the concepts and insights of → **causal learning**. For publications on causal machine learning, see for example [Kaddour et al., 2022](#), [Sanchez et al., 2022](#); [Lagemann et al., 2023](#); [Feuerriegel et al., 2024](#).

The causality-rooted approaches make causal machine learning a valuable tool in health and medicine, allowing for instance to model the effects of changes (“interventions”) on variables and what could have happened (“counterfactuals”) in hindsight. This enables researchers to investigate, in virtual model, outcomes resulting from a specific medical intervention or treatment and quantify these, allowing robust decisions even in the presence of confounding factors.

In addition, causal machine learning may be a promising avenue towards the goal of intelligible AI, interpretability or explainability of AI models (→ **interpretability and explainability**; → **interpretability**) and allow meaningful explanations of why a specific outcome was produced by an AI system, which is a challenge posed by “black-box” models (e.g. neural networks).

#### Explanatory note

While causal machine learning and → **neurosymbolic AI** have overlapping aims, their approach is different. Causal machine learning focuses on the modelling of robust cause-effect relationships in data as opposed to utilising rules-based symbolic information processing used by neurosymbolic AI. A key concern of causal machine learning is counterfactual reasoning, e.g. predictions of possible outcomes for alternative scenarios of a given process.

#### Term relationship:

Related terms:

- Causal learning
- Neurosymbolic AI

## Neurosymbolic AI

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

### Concept description

Neurosymbolic AI is an emerging concept in machine learning. The term refers to its two key constituting elements: 1) the use of → **artificial neural networks**, i.e. the use of statistical → **deep learning** techniques based on computational architectures inspired by neuronal cell ensembles, e.g. cortical columns in the mammalian cortex (hence the prefix ‘neuro’) and 2) rules-based symbolic information processing techniques (“symbolic AI”) based for instance on IF-THEN statements or decision trees. Neurosymbolic AI aims to combine explicit knowledge representation with the desirable properties of neural networks, namely their capacity for pattern detection in large data sets due to the distribution of statistical processing along many artificial neurons and their connections.

The principle aim of neurosymbolic AI is to enhance the → **intelligibility** (or explainability) of AI systems by encapsulating rules in the → **machine learning** process as well as enhancing their accuracy and precision. Neurosymbolic AI may be an approach to tackle the drawbacks of → **artificial neural networks**: their opacity and resulting inscrutability due to lack of → **intelligibility**, the occurrence of irrelevant, incorrect and unfaithful predictions (i.e. ‘hallucinations’ or ‘confabulations’) and (probably) related issues with recapitulating rules-based reasoning and mathematical operations, e.g. inability to compute simple addition or multiplication of large numbers, problems with counting tasks or multi-digit arithmetic calculations and overall inconsistency in regard to the application of mathematical rules (e.g. [Testolin, 2023; Satpute et al., 2024](#)).

### Explanatory note

Note that artificial neural networks do not reproduce brain architecture, let alone human brain architecture. Their basic design is inspired by cortical columns in primary visual cortex, largely based on neurobiological findings in cats, mice and macaque monkeys.

While neurosymbolic AI and → **causal machine learning** have overlapping aims, their approach is different. Neurosymbolic AI aims at incorporating human-like reasoning and explicit knowledge structures into neural networks.

### Term relationship:

Related terms:

- Causal learning
- Causal machine learning
- Deep learning
- Artificial neural networks

## Transfer learning

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

### Concept description

Transfer learning involves the use of a model that was pretrained on a large data set of one or more → **data modalities** that is subsequently further trained with a specific data set, tailored to the intended application. This task is typically aligned with the general → **AI typology** of the pretrained model (e.g. computer vision for pattern recognition in images). Transfer learning may also involve modifications of model architecture (in view of the required task, e.g. medical image analysis). For overviews of transfer learning in medical image analysis, see [Chuen-Kai S et al., 2015; Kora et al., 2022; Atasever S et al., 2022](#).

[2023; Salehi, 2023](#). The foundations of transfer learning were already elaborated in the 1970ies ([Bozilnovski, 2020](#)).

Transfer learning based on supervised models trained with a large set of natural images (e.g. ImageNet or iNat2021) is a frequently used technique for developing models for a variety of medical image analysis tasks, including image segmentation ([Isensee, 2021](#)). Adapting pretrained models to medical (e.g. radiological) images involves bridging the domain gap from natural to medical images ([Hosseinzadeh Taher MR et al., 2021](#)). Pretrained models (e.g. ResNet, VGG, DenseNet) can be obtained from various platforms. Examples of transfer learning for medical image analysis are numerous and include fundus image analysis for diagnosing diabetic retinopathy or analysis of pathology tissue slides for cancer detection.

Transfer learning has various benefits such as enhanced efficiency of the → algorithm-to-model transition, better performance (e.g. accuracy) and reduced cost. Transfer learning avoids developing an algorithm from scratch and builds on accuracy and adaptability benefits related to the large data set used for training the original model ([Hosseinzadeh Taher MR et al., 2021](#)). Transfer learning is thus useful in scenarios where training data are difficult to obtain and for tasks that align with a specific → AI typology. Models that were pretrained on fine-grained data, enable build performant tools through transfer learning with comparatively small datasets. Transfer learning may involve various techniques, e.g.

- fine tuning of model parameters (→ parameters) based on the new specific dataset. Fine-tuning can be achieved with considerably less data than training a ‘naïve’ model. Such fine-tuning may involve exploring the ideal → hyperparameters for the training process.
- feature extraction, i.e. the pretrained model is used to extract relevant data → features from the new training data which can then be used for, for instance, building a classification model.

Other techniques include domain adaptation, multitask learning, → federated learning & split learning, and few-/single-/zero-shot learning ([Chato & Regentova, 2023](#)).

## Explanatory note

### **Transfer learning: approaches**

An example of transfer learning in healthcare is the use of pretrained models for computer vision (e.g. “Resnet” networks (residual neural networks)) for developing quickly and with comparably little data need diagnostic image analysis models that are aimed at a specific clinical problem ([Chassaigne et al., 2024](#)).

The term ‘multistage transfer learning’ denotes specifically the transfer learning in medical imaging, by leveraging “pre-existing knowledge from a particular domain to enhance the performance of deep learning models in a target medical imaging domain” ([Ayana et al., 2024](#)).

Transfer learning can take various approaches, e.g. preserving the feature extraction layer of the pretrained model, while modifying the final (output) layers using the specific training data or using the pretrained model’s weights (between artificial neurons) as a starting point when training the neural network with new data.

### **Transfer learning and ‘fine-tuning’**

A specific type of transfer learning is ‘fine-tuning’ ([Microsoft, 2025](#); for an overview of fine-tuning of large-language models: [Parthasarathy et al., 2024](#)). Fine tuning seeks to preserve the ‘knowledge’ encapsulated in the pretrained model by updating or changing model parameters (→ parameters) in a stepwise and careful manner. Fine-tuning is particularly useful when the available data set for transfer learning is comparatively small and in cases where the pretrained model has properties that are highly relevant for the intended application of the final model. ResNet 50 for instance is a pretrained model for general image recognition tasks (“visual AI”; → AI typology) that can be fine-tuned for specific purposes, e.g. medical imaging data in view of diagnostic predictions or disease predictions.

## Term relationship:

Related terms:

- Machine learning
- Machine learning model

- Foundation models
- Generative AI
- Learning algorithm
- Algorithm-to-model transition

## Hybrid model / algorithm

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

### Concept description

A hybrid model approach draws on multiple approaches to achieve a new model that, ideally, would be superior to models achieved through one single approach (for an introduction to hybrid models, see [Sarker, 2022](#)).

### Explanatory note

Examples for a hybrid approach are the coupling of a fuzzy classifier module with an adaptive genetic algorithm ([Reddy et al. 2020a](#)), neuro-probabilistic hybrid approaches for improving diagnostic outcomes with AI ([Zhao et al. 2022](#)) or hybrid quantum convolutional networks ([Ajlouni et al. 2023](#)).

### Term relationship:

Related terms:

N.A.

## Hybrid learning

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

### Concept description

Hybrid learning is a → machine learning approach in the context mainly of generative → deep learning-based models that can learn from both labelled and unlabelled data. In such hybrid networks or models multiple deep basic learning models make up hybrid deep learning models, with the basic model being a discriminative or generative learning model (*concept description based mainly on Sarker, 2022*).

### Explanatory note

Generative models are versatile, learning from both labelled and unlabelled data. In contrast, discriminative models are unable to learn from unlabelled data yet may outperform their generative versions in supervised tasks. Hybrid networks are motivated by a paradigm for simultaneously training deep generative and discriminative models. Multiple (two or more) deep basic learning models make up hybrid deep learning models, with the basic model being the discriminative or generative deep learning model outlined previously.

### Term relationship:

Related terms:

- Deep learning
- Artificial neural networks

## Model development

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

### Concept description

Model development encompasses all steps towards building a model that is capable of generalising prediction or other relevant output when processing new, unseen samples of data. We consider model development part of the pathway of → algorithm-to-model transition.

In the case of → machine learning, model development is typically based on a dataset that is separated into three groups: → training data, → validation data, → testing data that are used for → model training, → model validation and → model testing. These are typically retrospective data: in case of AI systems used in medicine, these would be patient health data collected from hospitals, from past clinical studies, → clinical investigations or clinical trials or from relevant clinical registries.

Briefly, the typical approach of model development is as follows:

- *Model training:* Training of the → machine learning algorithm using the → training data to create a → machine learning model Keeping the training data separate from the validation and testing data is critical to avoid → overfitting.
- *Model validation:* initial assessment of → model performance with a separate → validation data set. Results from model validation are used to guide adjustments including parameters such as learning rates, regularisation to reduce → overfitting and fine-tuning of → hyperparameters.
- *Model testing:* final assessment of → model performance and → generalisability using an independent data set, the → testing data. Model testing is not sufficient for AI systems with a medical / clinical purpose. Additional evidence from → model evaluation and → clinical validation is required that assess a variety of aspects including performance under → real-world use conditions. These latter steps involve dedicated studies using prospective data (e.g. clinical studies, → clinical investigations, clinical trials).

Apart from choosing the AI type and algorithm, there are a number of the important decisions to be made during model development (see explanatory note) and it is generally advisable to conduct a → user research as early as possible and to get input from various stakeholders in view of requirements and constraints of both → use context and → use environment.

### Explanatory note

Data are a key element for developing a performant and robust AI model. But they are not sufficient (→ algorithm-to-model transition). Important decisions during model development include

- 1) Choosing the model architecture (or → AI technique, e.g. feedforward neural network, recurrent neural network, traditional ML approach such as support vector machine etc.). This decision will depend on, inter alia, the objectives, → intended use, intended → use context and → use environment, the → data modality/ies of → input data or the importance of → intelligibility or availability of appropriate explainability techniques.
- 2) The objective, i.e. the metric that the model is supposed to optimize, e.g. minimizing the mean squared error between its predictions and the actual labels (see → objective function: loss function, cost function) and
- 3) Other key elements, the optimizer (i.e. mathematical optimisation function) such as gradient descent or adaptive optimisers (e.g. RMSprop or Adam; see → objective function: optimisers)
- 4) Appropriate metrics for assessing the performance of the model, i.e. → evaluation metrics (e.g. recall, accuracy, precision for binary classification outputs; other metrics for multi-class predictions or non-classification outcomes).

After training, model parameters are adjusted or fine-tuned using the → validation data set before proceeding to → model testing / evaluation.

#### Term relationship:

Related terms:

- Machine learning model
- Algorithm-to-model transition (ATMT)

## Evaluation metrics

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

#### Concept description

Evaluation metrics are crucial for evaluating the performance of → machine learning models during

- the development phase in the pre-deployment space, e.g. during → model validation, → model testing or → model evaluation as well as
- in the post-deployment space, e.g. for continuous monitoring of → model performance, safety (→ AI safety), → bias. This concerns both, models that are ‘frozen’, i.e. do not learn during their use in the post-deployment space, as well as models that exhibit → continuous and adaptive learning, i.e. that are continuously trained on → post-deployment input data.

There are no universally applicable performance metrics, instead they must be carefully chosen based on several considerations, including → intended use, → use context and needs of end users. Importantly the consequences of potential errors (e.g. missed actual positives) must be considered. For instance, for → machine learning models used for clinical diagnostics or screening purposes, *recall* may be a metric that is more appropriate than *precision*. Evaluation metrics should be transparently documented to allow interpretation of the AI system’s performance characteristics by stakeholders.

Evaluation metrics provide quantitative measures that show whether specific decisions and actions (e.g. model selection, architecture decisions, fine-tuning of model parameters and adjustment of hyperparameters (→ parameters; → hyperparameters)) improve → model performance. Evaluation metrics

- allow to detect → overfitting, e.g. by comparing → model performance on → training data versus → validation data.
- support ‘early stopping’ by showing when → model performance improvements flatten (see also → environmental sustainability of AI system throughout the life cycle).
- facilitate comparisons between different approaches or models, thus accelerating → model development
- help determining whether a model is ready for more in depth evaluations (see → model evaluation) and / or deployment (e.g. assessing whether pre-defined minimum performance metrics are met)
- support the detection of → bias (addressing → FAIRNESS requirements)

Depending on the → AI typology, → AI technique and the output of the AI system, various evaluation metrics can be used.

- For **classification outputs providing discrete binary labels**, metrics include → accuracy, → precision, → recall or → receiver operating characteristic (ROC) curve. For multi-class classification problems, other metrics must be chosen ([Müller D et al. 2022](#)).

- For **regression outputs** (continuous range of values as output), metrics include: mean square error, root mean square error, mean absolute error, R-squared.
- For **language models**, metrics include
  - ‘Bleu/rouge scores’ that address lexical overlap (i.e. word-for-word matching with reference texts, without assessing semantics or truthfulness/factual correctness of generated outputs).
  - ‘Perplexity’, a measure of the uncertainty associated with the predictions of a language model, without however addressing → **usability** or truthfulness of outputs.

For additional metrics, see explanatory note.

## Explanatory note

### References and further reading

- A detailed overview over evaluation metrics can be found on the DeepAI machine learning glossary / definitions (DeepAI, 2025).

### **Evaluation metrics for medical ML applications**

- Hicks et al. (2022) provide a detailed and critical discussion evaluation metrics in medical ML applications mainly in gastroenterology. The review is however useful for various ML applications in medicine and healthcare.
- Müller D et al. (2022) provide considerations for evaluation metrics in the context of image segmentation.

### **Evaluation metrics for large language models (LLM) including LLM for healthcare**

A variety of other LLM evaluation metrics are available that try to address aspects such as fluency and text quality, faithfulness to the source texts, adequacy, human notions of correctness (e.g. BLEURT).

For a short overview on metrics see [Huang et al., 2024a](#); [Microsoft Ignite, 2025](#); [Shetty, 2025](#); [Github, 2025](#). We note that a controlled vocabulary for describing relevant attributes appears to be missing at present.

A recent review highlighted the need for better metrics for LLMs used in healthcare and identified issues with fairness, bias and ‘toxicity’ ([Bedi et al., 2025](#)).

## Term relationship:

### Related terms:

- Model validation
- Model testing
- Model evaluation
- Algorithm-to-model-transition (ATMT)
- Model performance
- Model calibration

## Model performance

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

### Concept description

Model performance concerns primarily the → accuracy of predictions or output based on → evaluation metrics, e.g. → accuracy, → precision, → recall, AUC-ROC (→ receiver operating characteristic (ROC)), mean square error (see → objective function). These metrics allow to assess whether a model is able to make sufficiently correct predictions on data it has *not* been trained on (see → model testing).

Importantly, a model can show impressive accuracy but provide insufficiently calibrated probability estimates (e.g. of disease). Thus, an important second aspect of assessing the usefulness of a model is → model calibration.

Focusing only on performance metrics (including calibration) may not sufficiently address issues of adopting AI models in healthcare since these metrics do not sufficiently reflect how a model / AI system would augment medical decision-making in a specific situation (see → model actionability).

### Term relationship:

Related terms:

- Algorithm-to-model-transition (ATMT)
- Model development
- Evaluation metrics
- Model performance
- Model actionability

## Model calibration

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

### Concept description

Calibration or '*calibration accuracy*' is an important additional aspect of → model performance in the context of predictive analytics, e.g. for clinical decision making. It refers to the degree to which a model is *calibrated* in regard to the → accuracy of actual risk estimates or probabilities (see [Van Calster B et al., 2016: "The accuracy of risk estimates, relating to the agreement between the estimated and observed number of events, is called calibration"](#)).

#### **Discrimination accuracy and calibration accuracy**

Performance indicators described under → model performance are typically indicators of *discrimination accuracy*. In contrast, model calibration accuracy captures the similarity between the predicted probability suggested by the AI system and the *real-world probability*.

This real-world or *actual* probability is influenced by the disease prevalence (or 'pre-test probability') as laid out in Bayes' theorem of probability ([Westbury, 2010](#)). Thus, an AI prediction model that does not take the pre-test probability into account may be inaccurate and misleading. Calibration accuracy needs to be addressed for AI systems in diagnostic contexts ([Park et al., 2021](#)); calibration is not always sufficiently considered by model developers ([van Calster et al., 2019](#)).

#### **Calibration as an estimation of the confidence of an AI model's predictions**

Calibration allows to estimate the *confidence* of a model's predictions: estimated risks can be unreliable even in case a model shows good discrimination. Despite the fact that calibration performance is recommended by the TRIPOD guidelines for prediction modelling studies ([Moons et al., 2015; see Annex I](#)),

studies and systematic reviews showed that calibration is less assessed than discrimination (see in [Van Calster et al., 2019; Park et al., 2018](#)). This may, depending on model use, lead to misleading predictions and thus potential harm in clinical decision making ([Van Calster & Vickers., 2015; Van Calster et al., 2019](#)).

#### Explanatory note

Model calibration may support the interpretability of a model, supporting confidence and trust in the model's outputs and allowing to compare various models with each other. Further, properly calibrated probabilities allow to measure the *uncertainty* associated with the model's outputs.

Note that model calibration can also be understood as dealing with *intrinsic incompatibilities* ([Char et al., 2020](#)) concerning test method characteristics and trade-offs between these characteristics (see → *intrinsic incompatibilities and 'trade-offs'*).

#### Term relationship:

Related terms:

- Algorithm-to-model-transition (ATMT)
- Model development
- Model performance
- Intrinsic incompatibilities and 'trade-offs'
- Evaluation metrics

## Model actionability

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

#### Concept description

Actionability refers to the capacity of a learning algorithm to transfer new knowledge to end users ([Kulesza et al., 2013, 2015](#); see also [Villone & Longo, 2021](#)). More specifically in a healthcare context, "model actionability" ([Ehrmann et al., 2023](#)) refers to the evidence available that an AI model or AI system will augment clinical decision making, specifically when compared to clinical judgement alone (see → *added value*).

Key characteristics such as → *model performance* and → *model calibration* on their own render valuable but insufficient evidence concerning the real-world added value of a model or AI system in context of clinical decision making (see also → *heuristics*).

## Causability

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

#### Concept description

Causability refers to the quality of explanations on how and why AI systems produce a given output. Causability has been proposed as a concept by Holzinger and colleagues ([Holzinger et al., 2019, 2021; Müller et al., 2022](#)), who argue that explainable AI (XAI) provides evidence on technical or 'mechanistic' explanations but that these are, on their own, insufficient for medicine.

Causability may be a useful concept for ensuring appropriate → intelligibility of AI models in contexts of high-risk use such as healthcare and vis-à-vis various users (namely healthcare professionals) and stakeholders, namely patients.

For AI systems or models that *do not* show intrinsic interpretability (these are so-called black-box models based on → deep learning), intelligibility entails techniques to explain the functioning and input-output logic of AI systems through explainability techniques ([Burkart & Huber, 2021](#)).

### Explanatory note

The causability concept overlaps to some extent with considerations of the quality and understandability of scientific explanations (see → [explanations of AI systems and their outcomes](#)). See also the review by Miller ([2019](#)) on explanation in AI and insights from the social sciences.

### Term relationship:

Related terms:

- Model actionability
- Intelligibility
- Interpretability and explainability
- Explainable AI (XAI)

## Accuracy

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

### Concept description

Proportion of correctly identified instances from all examined instances. Accuracy, for binary outcomes (negative, positive) of any predictive system or tool is equal to the sum of true positives + true negatives over the sum of true negatives + true positives + false negatives + false positives ([Griesinger et al., 2016](#); [Char et al., 2020](#)).

Accuracy of a → machine learning model is **influenced by various decisions made during → algorithm-to-model transition**, including inter alia → conceptual relevance, decisions impacting on → data quality, modelling decisions, including notably the choice of → objective function (loss function, cost function, other constraints and the optimiser).

Further, in → machine learning there is typically a **trade-off between accuracy** (or overall → model performance) and **interpretability or explainability** (→ interpretability and explainability): for instance, models with complex loss functions (→ objective function) may have higher accuracy – however at the expense of intrinsic interpretability, in particular in case a model is optimised for composite objectives (with multiple terms), which complicate disentangling which of these impacted a particular model output. Specific regularisation approaches (see → [overfitting](#)) in loss functions can support interpretability by highlighting the most relevant features.

### Explanatory note

Determining accuracy for a predictive AI system in healthcare is perhaps more complicated than in other applications areas. As pointed out by Char and colleagues ([Char et al., 2020](#)), medical diagnoses or decisions cannot be always labelled as clear-cut ‘incorrect’ or ‘correct’ due to various uncertainties and potential obfuscating factors, notably the selection of reference benchmarks (“gold standard”) against which predictive capacity is measured. This includes uncertainties of human rates, in case AI systems accuracy is measured against predictions by clinicians.

The difficulty of suitable “gold standards” is not a feature singular to healthcare: a similar problem exists in other areas of biomedicine, e.g. predictive *in vitro* test methods used in toxicology ([Griesinger et](#)

[al., 2016](#)). Further, due to these and other factors such as unknown complications to occur or, inversely, unexpected capacities for recovery, *downstream* → **health outcomes** are not always fully predictable.

Finally, the accuracy of an AI system will be influenced by real-world conditions such as the → **use context** and → **use environment**. Fully correct estimations of accuracy can only be obtained under real-world conditions, while accuracy measures under optimised conditions of a clinical investigation or clinical trial provide an indication of accuracy that needs to be used with caution.

## Sensitivity

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

### Concept description

We use the definition provided by Liu et al., (2022):

*“Proportion of actual positives that are correctly identified in a binary classification. It is also called the true positive rate (TPR), recall, or probability of detection.”*

### Explanatory note

N.A.

### Term relationship:

Synonyms:

- True positive rate
- Recall
- Probability of detection

## Specificity

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

### Concept description

We recommend the definition provided by Liu et al., (2022):

*“Proportion of actual negatives that are correctly identified in a binary classification. It is also called the true negative rate.”*

## Precision

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

### Concept description

We recommend the definition provided by Liu et al., (2022):

*“Proportion of predicted positives that are true positives.”*

### Term relationship:

Synonyms:

- Positive predictive value

## Receiver operating characteristic (ROC)

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

### Concept description

We recommend the definition provided by Liu et al., (2022):

*“A graph showing the sensitivity the true positive rate ( $TPR = \text{sensitivity}$ ) against the false positive rate. The area under the ROC curve is a measure for how well a specific parameter setting distinguishes between two groups”.*

## Precision-recall (PR) curve

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

### Concept description

We recommend the definition provided by Liu et al., (2022):

*“A graph showing the precision against the TPR (=sensitivity=recall) in order to display the trade-off between precision and recall for different parameter settings. The area under the PR curve is a measure for highly imbalanced classification tasks”.*

### Term relationship:

Related terms:

- Recall
- Sensitivity

## Recall

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

### Concept description

See → sensitivity

### Term relationship:

Synonyms:

- Sensitivity
- True positive rate

- Probability of detection

## True positive rate

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

### Concept description

See → sensitivity

### Term relationship:

Synonyms:

- Sensitivity
- Recall
- Probability of detection

## Replicability

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

### Concept description

Obtaining consistent results across acts of scientific inquiry (scientific studies, measurements, diagnostic tests etc.) aimed at answering the *same question*, each of which has obtained its own data and may have used different methods for collecting, measuring, compiling, processing and interpreting these data.

### Explanatory note

This elaboration of replicability (as well as → reproducibility and → generalisability) is mainly based on the excellent report of the US National Academy of Sciences on “Reproducibility and replicability in science (2009): “*We define reproducibility to mean computational reproducibility— obtaining consistent computational results using the same input data, computational steps, methods, code, and conditions of analysis; and replicability to mean obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data. In short, reproducibility involves the original data and code; replicability involves new data collection and similar methods used by previous studies. A third concept, generalisability, refers to the extent that results of a study apply in other contexts or populations that differ from the original one.<sup>1</sup> A single scientific study may entail one or more of these concepts.*”

See also Meng, 2020.

### Term relationship:

Related terms:

- Reproducibility
- Generalisability

## Reproducibility

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

### Concept description

Obtaining the same result over various acts of measurements, computations etc. (typically performed in one setting, e.g. laboratory).

For AI, reproducibility means that the AI systems produces the same outputs (computational results) when using the same input data and in situations where other potential variables have not been changed.

### Explanatory note

See explanatory note → [Replicability](#)

### Term relationship:

Related terms:

- [Replicability](#)
- [Generalisability](#)

## Reliability

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

### Concept description

#### ***Reliability – a function of reproducibility and replicability***

Reliability means that data, measurements, processes or scientific findings *can be relied on* since they are demonstrably *reproducible* and *replicable*: (→ [reproducibility](#) and → [replicability](#)). For traditional AI systems that produce discriminative outputs (as opposed to → [generative AI](#)), reliability means that the output of the system is reproducible and replicable and can therefore, from a perspective of repeated use, be relied on. Notably, reliability does *not* imply that the data, measurements etc. are necessarily *correct* – they may or may not be. However, in colloquial language reliability has a notion not only of reproducibility and replicability, but also of a certain relevance and usefulness. Further, for predictive test methods for example, reliability can be measured as the capacity of a test method to make *accurate* predictions as compared to reference data *over various measurements* within use environment (reproducibility) and across use environments (replicability). In such cases reliability and predictive capacity or accuracy are assessed as an ensemble.

In case of → [generative AI](#), it is important to keep in mind that outputs are probabilistic *data* predictions (as opposed to discriminative predictions of a ‘traditional’ AI model, e.g. ‘positive’ or ‘negative’ classifications). Thus, presenting for example an LLM with the exact same data and prompt repeatedly will lead to a genuinely new data prediction each time. Hence the outputs of the LLM over time may vary within and between environments. This is an intrinsic feature of generative AI. Thus, reliability and reliability measurement is associated with more uncertainty.

#### ***Assuring reliability through dedicated evaluation and validation***

Reliability is a key requirement for routine use of AI systems in healthcare and other health applications. Sufficient effort is required to show reliability under varying use conditions, e.g. → [use environments](#). This can be done to some extent by → [model validation](#), but requires moreover processes including → [model evaluation](#) and → [clinical validation](#).

#### ***Reliability and uncertainty***

Reliability is related to the important scientific concept of *uncertainty*, i.e. the notion that all data, measurements, assumptions and scientific knowledge is associated with potential epistemological pitfalls, e.g. due to limitations of measurement techniques, imprecisions, the need to extrapolate, base assumptions etc.

In the case of AI systems in health, uncertainties may stem from a variety of sources such as invalid assumptions, outdated scientific theories (e.g. about disease origin), biased data, incomplete data and also fundamental limitations such as Gödel's incompleteness theorems (Fleetwood, 2023). Thus, even where reliable findings agree with prevalent theories or scientific consensus, there is no absolute guarantee of truthfulness.

#### Explanatory note

See explanatory note → [Replicability](#)

#### Term relationship:

Related terms:

- [Replicability](#)
- [Reproducibility](#)
- [Generalisability](#)

## Generalisability

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

#### Concept description

The extent to which results, → outputs and output data of an AI system are applicable to another than the original context (→ contextual relevance) and hence 'generalisable'.

For example, if an AI system used for diagnostics is found to be also applicable to other diseases and conditions than originally designed for.

Generalisability in healthcare entails applicability to more varied patient populations (see → [model evaluation](#)) in varying → [use environments](#).

#### Explanatory note

See explanatory note → [Replicability](#)

#### Term relationship:

Related terms:

- [Replicability](#)
- [Reproducibility](#)

## Overfitting

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

#### Concept description

Overfitting is a phenomenon associated with the extensive use of a typically small → training data set which results in the model matching the training data to such an extent that it fails to produce correct

outcomes (e.g. predictions) on data outside the training data domain. Overfitting leads to poor generalisability of models (Liu et al., 2022).

Overfitting is especially an issue of → machine learning algorithms that exhibit a high degree of dimensionality and mathematical complexity (e.g. → artificial neural networks used for ‘→ deep learning’).

These models have a strong data dependence and may show seemingly convincing → model performance (e.g. high → accuracy) in training data, but show reduced performance when assessed with → validation data and in particular *external validation data* (→ model validation; → model evaluation; → clinical evaluation) (Park et al., 2021).

While overfitting can be reduced by various techniques, collectively referred to as ‘regularisation’, these approaches are not always successful, leading to reduced → generalisability of AI models, in particular in the area of diagnostic prediction (Park et al., 2021).

### Explanatory note

A summary of overfitting in a clinical context can be found in Mutasa et al., (2021) and Schinkel et al., (2019).

An example of a simple approach to tackle overfitting in → artificial neural networks is ‘dropout’: this involves the random disablement of artificial neurons during training, reducing the impact of specific neurons and hence their weights (→ deep learning) and bias (→ neural network bias) on the networks overall output (Srivastava et al., 2014).

### Term relationship:

Related terms:

- Model development
- Development data
- Training data
- Validation data
- Testing data

## Foundation models

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

### Concept description

The term “*foundation model*” (FM) was introduced by Bommasani et al., in 2021. FMs are AI models based on → deep learning in → artificial neural networks. The defining characteristic of FMs is that they are trained with a particularly large amount of general and unlabelled data, rendering them potentially useful in a wide spectrum of → use contexts and for various applications or tasks: the models are hence “foundational”.

FMs represent a shift from task-specific AI (e.g. an AI system able to predict lung cancer from medical images) to AI that is able to address various tasks without a need for substantial retraining. Thus, FMs are also a novel way of developing AI in a flexible manner that avoids building an AI model for specific application from scratch.

Foundation models are often based on a so-called “transformer” neural network architecture (see → artificial neural network – types; → deep learning), entailing a technique called ‘attention’ to allow the model to learn complex and ambiguous relationships e.g. between words, a typical feature of human language (see the seminal paper on ‘attention’ by Vaswani et al., 2017).

### Mono- and multimodality of FMs

FM can be *mono-modal*, i.e. trained on one → data modality, e.g. text (e.g. large language models, LLM and associated ‘chatbots’ that create textual / language output; see- > conversational agent). FMs can

also be *multi-modal*, i.e. trained with data of various modalities. WHO has issued in 2024 guidance on large multi-modal models for AI in health ([WHO, 2024a](#)).

### FMs and generative AI: close relatives

FMs can also be seen from the perspective of their capacity to “generate” (or rather *synthesize*) new data that are based on the vast body of training data used for developing the FM model. The property of synthesizing content is called “generative AI” (→ [generative AI](#)). While the term ‘foundation model’ focuses on the broad training of the model and its applicability to a wide set of tasks (i.e. its ‘foundational’ character), the term generative AI focuses on the data output type of FMs, i.e. new synthesized or generated data or so-called ‘content’, to emphasize the immediate meaningfulness of such data for humans.

### Explanatory note

- For a short discussion of FMs in medicine and healthcare, see → [generative AI](#).
- For an excellent introduction to FM, see UK Competition and Markets Authority ([2023](#)).
- Stanford University runs a "Center for Research on Foundation Models (CRFM)" dedicated to exploring both technologies and societal consequences of FMs. Online:  
<https://hai.stanford.edu/news/introducing-center-research-foundation-models-crfm>

### Examples of FMs

Examples of mono-modal FM include

- models trained with text data (large language models, LLM) such as ‘ChatGPT’, ‘Mistral AI’, ‘BERT’, ‘Claude AI’ or ‘DeepSeek’
- visual AI models such as ‘DALL-E’
- auditory data trained with auditory signals including music (e.g. ‘MusicGen’).

LLMs can also cross modality borders: ‘MusicLM’ for instance generates music that tries to reflect textual prompts ([Agostinelli et al., 2023](#)).

All these models are ‘generative’ (→ [generative AI](#)), i.e. able to synthesize new data or content based on the training data on which the neural network was run on. LLMs have also been adapted to genomic data ([Zvyagin et al., 2023](#)).

### Term relationship:

Related terms:

- Alignment
- Generative AI
- Data modality
- Generalisability
- Artificial neural networks
- Deep learning
- Large language model (LLM)

## Generative AI (GenAI)

### Cluster: B.8 Algorithm, model, algorithm-to-model transition

### Concept description

The term ‘generative AI’ refers to → [machine learning models](#) that feature the property of being able to synthetise (or ‘generate’) new data or ‘content’, such as text, syntax or symbols (LLMs), images, sounds

(GANs, diffusion models), algorithmic code, structures of chemicals etc. The term 'content' means data with contextual meaning for humans. In a health context, content could refer to a medical report compiled based on text notes by a clinician or recorded conversations, a concise summary of clinical literature or a synthetic medical image (→ synthetic data). This data generation is based on *predictions* generated by a trained model mapping the underlying data distribution of the training data themselves: for example in case of LLMs, the model predicts the next most likely word ('next token prediction loss'; → objective function), the next most plausible pixel (per-pixel loss functions) etc. and hence 'generates' what human perceive as content. It is critical to bear this in mind when using generative AI: it is ultimately a probabilistic 'guessing' machine.

Generative AI is typically but not necessarily a feature of → foundation models which are not bound to one task. Generative AI thus differs from 'classical' machine learning models that typically provide 'discriminative' outputs (→ output and output data) focused on one specific task such as a diagnostic or prognostic prediction for a specific disease (e.g. classification outcome). Generative AI models are based on → deep learning in → artificial neural networks and are, given their architecture and large training data set, typically → foundation models. The → data modality/ies of the output of generative AI depends on the data on which these models were trained. Multi-modal generative AI models can predict data of different types. A summary on key application areas of GenAI and challenges concerning its implementation in health can be found in [Ceresa et al. \(2025\)](#).

#### **Data quality and generative AI can relate to each other in several ways**

- Generative AI *a*) may require foundational data (the huge amount of data trained to analyse patterns, structures, forecast likely linkages and generate (or synthesize) context, e.g. LLMs) and *b*) contextual data that supplement the training data with additional cues for generating fitting prompts ([Roch, 2024](#)). Both data types need to be of high quality, given the multiplier effect of foundation models / generative AI. Yet, realistically, issues of data quality, data veracity, data "toxicity" as well as issues with privacy have not been fully addressed in the huge training sets. Thus, "aligning" these models with values such as faithfulness, truthfulness and ensuring that they are of utility to users is a key aspect of generative AI (see → alignment).
- Generative AI (→ foundation models and generative AI) may be used to improve data quality, e.g. by creating → synthetic health data, thus augmenting existing datasets, by cleansing data sets (identifying and correcting errors and anomalies, including duplications, missing values etc.) or by enriching data with new → features / → attributes ([Xenonstack, 2023; Confluent, 2024](#)).
- Finally, the increasing dissemination on the internet of AI-generated data with no real-world relevance and no true veracity is a concern for data quality. If such data are used uncritically for training specific AI systems (e.g.), this might negatively impact the quality of a → model or → AI system (see explanatory note and reference under → output / output data).

#### **Explanatory note**

##### **FMs and generative AI in medicine and healthcare**

Existing FMs (mostly LLMs) that have the property to generate data (generative AI) are being critically discussed as tools also for **medicine and healthcare** (e.g. [Clusmann et al., 2023; Thirunavukarasu et al., 2023; Sedaghat et al., 2023; Wornow et al., 2024; Howell, 2024; Reddy, 2024](#)). Uses include assisting with the composition and/or analysis of patient records (e.g. [Yang et al., 2022](#)) or as highly accurate predictors, including in health systems ([Jiang et al., 2023](#)), to analyse big health data (e.g. emerging European Health Data Space, EHDS) or to support hypothesis-formation and knowledge creation. While FM have considerable potential, there are obvious risks.

What regards healthcare uses of foundation models or generative AI, there appear to be three areas of concern, with numbers 2 and 3 more specific to these models:

**1) Bias and equity:** GenAI may propagate inadequacies ([Yang et al. 2024](#)) hidden in the training data. These include historical biases (race, gender, social status) or outdated medical concepts, encapsulated in scientific and clinical publications. Such conceptual inadequacies (→ conceptual relevance) and biases

might lead to diagnostic errors, inequitable treatment recommendations, and further health disparities. Addressing bias requires conscious efforts in data collection, algorithmic design, model auditing, and deploying fairness-aware machine learning techniques throughout the GenAI lifecycle (see → [algorithm-to-model transition](#)). Although this risk is not unique to GenAI, the potentially future role of GenAI cutting through many health aspects poses a particular risk. Notably, most models are currently either trained on too narrow a dataset or broad public corpora (notably PubMed) and are evaluated on tasks that do now allow gauging their real-world usefulness for health systems ([Wornow et al., 2023](#)).

**2) Incorrect content:** the risk of generating outputs that seem, *prima facie*, plausible but are, on closer inspection, nonsensical or not rooted in true epistemic data ([Rawte et al., 2023](#)). Such contents are called 'hallucinations' or 'confabulations'. Sun and colleagues have proposed a classification system for such distorted content ([Sun et al., 2024](#)). Although the outcomes, i.e. their factual relevance, may be improved through → [alignment](#) and other approaches ([Roit et al., 2023; Ceresa et al., 2024](#)), there remains a high degree of uncertainty: results may not be relevant or, worse, simply 'made-up' (based on probabilistic guesses). Thus, great caution is required when employing FMs in medicine and healthcare.

**3) Echo chamber of probabilistic processes:** 'Contents' created by GenAI are ultimately rooted in probabilistic processes of data representations and learned 'semantic contexts' incorporated in billions of parameters within artificial neural networks. There is the risk that the output is to some extent a mere sophisticated 'echo chamber' of the data selected to train the model ([Wornow et al., 2023](#)) and their stochastic connections shaped during machine learning. Paired with automation bias and complacency ([Arnold, 2021](#); see → [avoiding automation bias](#); → [avoiding automation complacency](#)), this could devalue human medical expertise (→ [deskilling](#)) and creativity (see however [Sæbø & Brovold, 2024](#)), leading to the propagation of care models rooted in specific data and algorithms. It might also impact on human dimension in medical care (→ [upholding a trustful patient-physician relationship](#)).

Using generative AI in healthcare will require further research but also precautions, including:

- **awareness of their limitations** (in particular that these models are not databases, able to retrieve information but essentially stochastic machines predicting outcomes based on probabilities derived from the training data),
- **appropriate user training** (e.g. the importance of prompting in regard to output quality) and
- **guidelines for their use** need to be established prior to deploying them for routine (clinical) applications. This might include recommendations regarding use exclusions. Soft guidance by medical association and research bodies or consortia might support the responsible use of generative AI ([Wornow et al., 2023](#)).

Notably, in addition to considerations of using existing FM, there are also projects to develop genuine biomedical foundation models ([Tu et al., 2023](#)). This might enhance the use of generative AI in health.

#### Term relationship:

Related terms:

- [Alignment](#)
- [Foundation models](#)
- [Data modality](#)
- [Generalisability](#)
- [Artificial neural networks](#)
- [Deep learning](#)
- Large language model (LLM)

## Embeddings

**Cluster:** B.8 Algorithm, model, algorithm-to-model transition

### Concept description

Embeddings are a way of converting complex input data into a form that can be processed by a → **artificial neural network** with hidden layers, e.g. a large language model (LLM). Embeddings help 'encoding' salient features for processing in subsequent layers. Technically, embeddings are vector representations of objects in such models, mapping high-dimensional data into vector spaces. Embeddings are typically obtained by algorithms that are called 'encoders'. These are designed to capture important features in data, to discard noise and to reduce dimensionality. In the context of LLMs, embeddings represent the semantic relationships between 'tokens', i.e. words, parts of words or any other relevant units of text representing human language. Embeddings are optimized during learning and by means of the → **objective function**.

### Explanatory note

For an introduction of embeddings in the health context, see [Howell, 2024](#).

### Term relationship:

Synonyms:

- Foundation models
- Generative AI
- Artificial neural networks
- Artificial neural networks - types
- Deep learning

## B.9 Relevance

### Conceptual relevance

**Cluster:** B.9 Relevance

#### Concept description

##### **Basic understanding**

With the term conceptual relevance, we capture the extent to which conceptual foundations (e.g. scientific, technical or clinical assumptions, knowledge and concepts) that underpin the design, specifications and output of an intended AI system are indeed relevant for the specific problem to be solved by the AI system. We consider decisions and evidence concerning conceptual relevance part of the → algorithm-to-model transition.

##### **Meaning of ‘relevance’**

Relevance here relates to the plausibility and meaningfulness of these foundations in view of addressing the real-world problem. Such relevance allows to gauge the likeliness to which an AI system may successfully contribute to solving the problem.

Conceptual relevance can for instance relate to the available evidence on the relationship between the occurrence of a specific biomarker or an altered anatomical structure in relation to a disease, or the significance of a specific epitope for drug binding in the context of health research and drug development.

An AI model that is built on such foundational knowledge which is backed up by robust evidence is *a priori* more trustworthy and likely to contribute to providing relevant information in relation a real-world problem.

##### **Lack of conceptual relevance in data features**

While the emphasis of ‘conceptual relevance’ is on human knowledge and understanding, conceptual relevance also applies to specific data features which may or may not have such relevance. → Deep learning models (see also → artificial neural networks) may emphasise irrelevant features with no conceptual relevance (see → shortcut learning).

#### Explanatory note

In the area of medical devices (MDs) and in vitro diagnostic medical devices (IVDs), the terms “**valid clinical association**” (for MDs) and “**scientific validity**” (for IVDs) are used to capture their conceptual relevance. These terms are of use also for → AI-enabled medical device software ([IMDRF, 2017a; EU MDCG, 2020](#)).

N.B. the term ‘conceptual relevance’, albeit with a slightly different meaning, is also used in relation to data mining and semantic searches by AI (e.g. [Ibrahim MH et al., 2021](#); [Tyler, 2024](#)).

#### Term relationship:

Related terms:

- Valid clinical association / scientific validity
- Contextual relevance

## Contextual relevance

### Cluster: B.9 Relevance

#### Concept description

With contextual relevance we refer to the relationship between an AI solution and a specific → use context (e.g. a clinical workflow, clinical pathway). Contextual relevance addresses the question whether and how the specific AI solution might provide added value to the → use context. Contextual relevance relates to → model actionability.

The contextual relevance should be well considered before deciding on design specifications of an AI system and, in particular, when deploying an AI system for another context than the one it was designed for.

#### Term relationship:

Related terms:

- Conceptual relevance
- Valid clinical association / scientific validity
- Model actionability

## Valid clinical association / scientific validity

### Cluster: B.9 Relevance

#### Concept description

According to the [WHO \(2021\)](#), valid clinical association answers the question whether there is a plausible scientific explanation for a device's use-case.

The IMDRF' definition ([IMDRF, 2017a](#)) is slightly narrower: valid clinical association means the association of the output of an AI system with a clinical condition or physiological state (i.e. not necessarily supporting the use case). See also the MDCG guidance on clinical evaluation for information on valid clinical association ([EU MDCG, 2020](#)).

Most importantly, a valid clinical association should support that there are sound scientific principles underpinning the use of the AI system (→ conceptual relevance). Depending on the situation, establishing a valid clinical association before using a device may prove difficult. In such cases → clinical performance data can serve as an additional input to support an assumed clinical association (produced by → clinical evaluation activities, e.g. → clinical investigations or → post-market surveillance).

Clinical performance data may be important in cases of the AI-enabled device being used in another → use contexts than the one it was designed for (e.g. off-label use of devices for orphan problems, e.g. diagnostics of rare diseases) – see also → contextual relevance.

Evidence supporting valid clinical association can be generated e.g. through literature research, professional guidelines, proof of concept studies, or manufacturer's own clinical investigations/clinical performance studies.

#### Explanatory note

According to the IMDRF, a valid clinical association is an indicator of the level of clinical acceptance and how much meaning and confidence can be assigned to the clinical significance of the AI system's output in the intended healthcare situation and the clinical condition/physiological state.

### Term relationship:

Related terms:

- Conceptual relevance
- Contextual relevance
- Clinical performance
- Clinical evaluation
- Clinical investigation

## B.10 Verification, validation, evaluation

### Verification

**Cluster:** B.10 Verification, validation, evaluation

#### Concept description

In line with the general understanding of the term, verification of an AI system comprises a process of activities aimed at verifying that the AI system a) as a whole as well as b) its constituting components or parts (→ AI system component / part) correspond to and meet earlier established design specifications.

Design specifications are chosen on the basis of the → intended use and may be based on dedicated → user research to understand needs and conditions of → use context and → use environment. They may already take usability aspects into account (→ usability validation). Design specifications are typically developed and approved during the design stage and, where needed, further refined during the development stage of the → life cycle of AI in health. Importantly, verification should examine and verify that the AI system properly functions properly within the range of specifications of a defined → use environment in which the system is supposed to operate, e.g. testing of e.g. a data exchange protocol of an AI system connected and interacting with other devices.

#### Explanatory note

The scope of verification of AI systems encompass the system as designed and may draw on previous verification of individual components of the system (e.g. hardware, user interfaces, actuators, sensors etc.).

The IEEE standard on system, software and hardware verification and validation (IEEE, 2016) provides an approach that is useful in the present context. It outlines relevant verification and validation processes and explains their relationship to the life cycle of a product.

According to this document, verification can be seen as: (A) The process of evaluating a system or component to determine whether the products of a given development phase satisfy the conditions imposed at the start of that phase. (B) The process of providing objective evidence that the system, software, or hardware and its associated products conform to requirement specifications (e.g., for correctness, completeness, consistency, and accuracy) for all activities during each life cycle process; satisfy standards, practices, and conventions during life cycle processes; and successfully complete each life cycle activity and satisfy all the criteria for initiating succeeding life cycle activities.

#### Term relationship:

Related terms:

- Validation

### Validation

**Cluster:** B.10 Verification, validation, evaluation

#### Concept description

Generally, validation is the process through which a given item, product, process or service that is being developed or has been obtained for a specific context (e.g. a medical device, computer programme, production process, test method, methodology or measurement) is judged to be "valid" (suitable, useful,

meaningful, reliable, informative etc.) in that given context (→ use context) and/or for a specific → intended use.

In the current context, validation can address and relate to various aspects of an AI system, namely

- → Model validation. This may include checking the integrity of the model → model integrity checking. Both → model testing and → model evaluation are closely related concepts. Model evaluation is used in the context of medical / clinical AI applications.
- → AI system validation (also referred to as ‘technical or analytical validation’).
- → Usability validation of the AI system
- → Clinical validation (a requirement for AI systems used for medical purposes). A closely related concept is ‘evaluation of digital health interventions / AI-SaMD’ (see [WHO, 2021b](#)). Both may entail dedicated clinical studies, clinical trials and → clinical investigations.

Validation (and verification) processes will happen at a dedicated stage of the life cycle (→ life cycle of AI in health). → Model integrity checking should moreover be performed at regular intervals once the model / AI system has been deployed.

#### Explanatory note (at present copy / paste)

Notably, in the context of AI-enabled medical device software, the concepts of → AI system validation, → clinical validation (typically including → usability validation) are typically considered part of → clinical evaluation, i.e. the ongoing process of evaluating the clinical safety and performance of a product.

Clinical evaluation extends from the pre- to the post-deployment/market phase. It also includes → post-market surveillance (e.g. for AI-enabled medical devices) and → post-deployment monitoring (for other AI systems not used in health but not healthcare).

The WHO has published an excellent report on training, validation and evaluation of AI-based medical devices ([WHO, 2021](#)).

For a general guideline on system, software and hardware validation and verification, see IEEE ([2016](#))

#### Term relationship:

Related terms:

- Model validation
- Validation data
- AI system validation
- Usability validation
- Clinical validation
- Clinical investigation

## Model validation

**Cluster:** B.10 Verification, validation, evaluation

#### Concept description

Model validation is a broad term that, on the most general level refers to an initial evaluation of a model’s ‘quality’ ([Google machine-learning glossary](#)) and validity for a given purpose (e.g. by evaluating its → accuracy or → model performance).

- Model validation sheds light on whether the model, typically a reduced and simplified representation of the real-world problem (see also → proxies), recapitulates this problem to a sufficient degree as to be of *likely* practical value.

- Model validation usually serves to fine tune models (the validation data set is sometimes also referred to as ‘holdout set’ or ‘tuning set’ (Park et al., 2021). Results obtained during model validation can guide adjustments and improvements of the model (‘fine-tuning’), for instance learning rates (→ hyperparameters), modification of network architecture, regularisation to reduce → overfitting.
- Model validation is an important step before validating the entire AI system → AI system validation and its usability (→ usability validation). In case of AI systems with a medical purpose (e.g. a clinical decision support AI system with a dedicated user interface), other validation steps are required in order to assess and confirm the performance, efficacy, efficiency, safety and usability of the AI system (→ model evaluation; → clinical validation).
- Model validation is usually performed using the → validation data set, typically a dedicated slice of the → development data set. Since the → validation data set differs from the → training data set, validation helps to guard against → overfitting a model (i.e. aligning it too much to a narrow training set, which compromises applicability of the model to other data, including from real-world settings, thus compromising → generalisability).
- Model validation usually precedes → model testing, i.e. evaluating the model against, typically, a third slice of the development data the → testing data.
- In case of models/AI systems intended for healthcare applications, elements of model validation and → model testing may be followed up by → model evaluation (Vokinger et al., 2020). Model evaluation focuses on → generalisability of a model to independent groups of patients, but also detection and avoidance of bias, errors etc. In contrast to model validation and testing which are conducted normally exclusively on retrospective data, → model evaluation may entail prospective data generated in the context of clinical studies, clinical trials and → clinical investigations.
- Experiences from model validation, → model testing and → model evaluation should feed into → clinical validation and → clinical evaluation, in particular in regard to clinical utility, → generalisability to broader patient groups and → real-world use and real-world → use environments.

## **1 Internal, external and local validation**

*Internal validation:* Model validation is typically first addressed by internal validation, using a subset of the → development data set, the so-called → validation data (also referred to as ‘holdout set’) which, importantly, has not been utilised for model training. Evaluating the model against → validation data may lead to subsequent fine-tuning of model → parameters and help with → model calibration.

*External validation* refers to the use of data from a different source to the internal → validation data set or, more generally, the → development data set. External validation allows examining to which extent the → model performance is generalisable to other scenarios, including clinically relevant scenarios in case of AI systems for healthcare (WHO, 2021b; UK NHS, 2024).

For AI systems used in healthcare, the → validation data must correspond to patient data. These may be enriched by → synthetic health data. For internal validation exercises, patient data are typically *retrospective*, i.e. derived from past studies or datasets, including clinical trials where available and useful.

External validation may be based on data from real-world patient cohorts and allow examining → usability aspects. It may also allow fine-tuning specific aspects of the model, e.g. in case biases were detected. While also external validation typically uses retrospective data, early use of prospective data (i.e. data generated in a specific healthcare setting collected in real-time), is highly desirable (see → model evaluation).

[Local validation](#): Depending on how and where the AI system will be deployed, external validation may focus on specifics of a given location, e.g. a local healthcare setting that may be characterised by specific patient characteristics in order to evaluate whether the → model performance adequately meets locally specific requirements.

## **2 Validation of performance**

In case of a predictive model with binary outcomes, model validation will entail an assessment of key performance parameters such as → accuracy, → sensitivity, → specificity, → receiver operating characteristic (ROC), → recall etc (→ model performance). For models that have multi-class prediction models or that predict values (rather than classes/categories), other methods exist. Carefully selected benchmarks can also be used for model valuation. Importantly, diagnostic prediction models should be calibrated for the so-called pre-test probability (see → model calibration).

## **3 Model validation and integrity verification or validation of the model**

Model validation may incorporate also → model integrity checking.

### **Explanatory note**

#### **General model validation approaches**

The → validation data set is one of the three subsets of the → development data set (the others being the → training data and the → testing data).

Typically, the trained model is evaluated against the validation set several times before evaluating the model against the test set. Well-established approaches include ([Matheny et al., 2022](#)):

- *K-fold cross validation: A dataset is randomly partitioned into K parts and one part is set for testing, and the model is trained on the remaining K-1 parts, and the model is evaluated on the so-called “holdout” part (validation data set).*
- *External cross validation: Perform cross validation across various settings of model parameters and report the best result.*
- *Internal cross validation: Perform cross validation on the training data and train a model on the best set of parameters*

### **Guidance on evaluation and validation of AI technologies (models/systems)**

Useful guidance on evaluation and validation of AI technologies in health can be found in:

- [World Health Organisation \(WHO\) \(2021b\) Generating Evidence for Artificial Intelligence Based Medical Devices: A Framework for Training, Validation and Evaluation.](#)
- [UK National Health Service \(NHS\) 2024\) Chapter 3.2 evaluation and validation. In: Understanding healthcare workers' confidence in artificial intelligence \(AI\) \(website training resource\).](#)

For a general approach to the validation of models (in particular for models where there is uncertainty concerning objects whose variables, parameters or scales are not fully controlled), please see:

- [Sornette D et al. \(2007\) Algorithm for model validation: theory and applications.](#)

### **Term relationship:**

Related terms:

- [Model testing](#)
- [Model evaluation](#)
- [Model integrity checking](#)

## Model testing

**Cluster:** B.10 Verification, validation, evaluation

### Concept description

Model testing is typically the last step of → model development (see also → algorithm-to-model transition). Model testing is performed using the → testing data set, a completely separate test dataset from → training data and → validation data. Model testing normally makes use of defined → evaluation metrics or suitable benchmarks.

Data obtained during model testing provides a final performance assessment of the model under 'laboratory conditions'. Model testing provides an estimation of the model's potential → generalisability. Importantly, data from model testing are not sufficient for the uptake of AI systems healthcare workflows: additional rigorous testing under real-world conditions and using prospective health data are required, see → model evaluation, → clinical validation, → clinical evaluation.

### Explanatory note

Importantly, **model testing does not entail model tuning** (i.e. optimisation, improvement) of the model. Using data obtained during model testing for further model would lead to overly optimistic performance estimates.

In contrast during → model validation (the step preceding model testing) the model is typically tuned: the → validation data set is traditionally also called 'tuning set' (Park et al., 2021).

### Term relationship:

Related terms:

- Model validation
- Model evaluation
- Algorithm-to-model transition (ATMT)

## Model evaluation

**Cluster:** B.10 Verification, validation, evaluation

### Concept description

The term 'model evaluation' has been proposed to denote the assessment and subsequent improvement of models intended for healthcare applications (Vokinger et al., 2021). The concept is linked to the concepts of → model validation and → model testing.

One of the main aims of model evaluation is to assess how well the model makes predictions in independent (sub)groups of patients and in independent clinical studies and/or trials (Vokinger et al., 2021; Bajwa et al., 2021). From this perspective, model evaluation may feed into → clinical validation and → clinical evaluation.

Model evaluation provides an initial 'validation' of the → generalisability of the model to different patients and the model's performance under more realistic conditions aligned with the → intended use, → use context and → use environment. Model evaluation may help detecting errors and → bias.

Like → model validation, model evaluation may incorporate 'post-processing' steps in view of improving the model prior to routine use or prior to → clinical investigation, → clinical validation and deployment.

In summary, model evaluation and post-processing can entail the following aspects:

- refining model outputs (e.g. threshold setting and refinement; see explanatory note)

- improving → model performance
- → generalisability to a wider group of patients
- addressing errors and reducing and mitigating → bias
- preparing the model for real-world application, with a view to considerations of → usability, → use context and integration of a model / AI system into a (clinical) workflow as well as aspects linked to variability of → use environments.
- Interrogating and improving the model's the → intelligibility, → interpretability and explainability.

Given the possible use of prospective data (e.g. during trials), model evaluation studies can contribute to → clinical validation of an → AI-enabled medical device software and therefore contribute to the overall evidence generated in the context of → clinical evaluation, a process that extends from the pre- to the post-deployment space.

The information gathered within restricted model validation exercises may help with the planning of → clinical investigations and help rectifying errors and reducing risks for trial participants before embarking on more comprehensive studies.

The proposed → DECIDE-AI tool goes into a similar direction by aiming at assessing performance, safety, usability in small clinical settings prior to routine deployment of an AI solutions.

#### Explanatory note

An example of a *post-processing* step is appropriate threshold setting: A model that provides a probabilistic prediction (from 0 to 1) for patient to develop a specific disease may require, for reasons of → usability in → clinical practice, to be adapted to provide a classification output (e.g. 'likely', 'un-likely', 'inconclusive'). In that example, post-processing would include refining and/or identifying and setting a threshold for translating probabilistic values into classification outcomes that is suitable for the intended use.

#### Term relationship:

Related terms:

- Algorithm-to-model-transition (ATMT)
- Model validation
- Model testing
- Model calibration
- Intended use
- Usability
- Intelligibility
- Interpretability and explainability
- Clinical validation
- Clinical evaluation
- DECIDE-AI

## Model integrity checking

**Cluster:** B.10 Verification, validation, evaluation

#### Concept description

Model integrity checking, also referred to as model integrity verification or validation) encompasses all actions seeking to establish that the model has not been compromised by malicious attacks (e.g. during

training or post-deployment), fulfils all required design specifications and performs as intended, without its integrity compromised in any form.

Model integrity checking is critical in particular for large mono- or multi-modal models (→ foundation models; → generative AI). Model integrity checking can be part of → model validation.

Various techniques have been established ([Nightfall AI, 2024](#)), including checksums and hashing, provenance tracking, runtime behaviour analysis and watermarking ([Hoque et al., 2023](#)).

### Explanatory note

Model integrity checking should not be confused with other necessary checks, e.g. data integrity checks that aim at affirming the validity, plausibility and relevance of data within a given healthcare or medical context.

### Term relationship:

Related terms:

- Verification
- Model validation
- AI system validation

## AI system validation ('analytical / technical validation')

### Cluster: B.10 Verification, validation, evaluation

### Concept description

AI system validation entails the validation of the AI system's input-output performance (e.g. the → accuracy of its diagnostic prediction model). Software validation exercises of this type can also be referred to as 'analytical' or 'technical' validation ([Goldsack et al. 2020](#)). (see for instance [IMDRF, 2017a](#) or).

For complex AI system, AI system validation may address the AI system in its totality, e.g. the → machine learning model as well as interplay between all → AI system components / parts (e.g. user interface, sensors or actuators of an AI-augmented robotic surgery suite). While individual components may have been validated on their own, individual validation records may not be sufficient for validating the system due to insufficient evidence on the correct interoperability of components.

AI system validation does not cover the validation of an AI system's → clinical performance and → clinical effectiveness which is done successively through → clinical validation and/or other approaches. AI system validation generally will emphasize aspects of technical performance, such as sensitivity and specificity and overall functioning under realistic conditions, either within research settings or real-world environments.

### Explanatory note

It should be noted that valuable relevant guidance documents and standards, for instance on software validation for medical devices ([US FDA, 2002](#)), on system, software and hardware verification and validation ([IEEE, 2016](#)), on clinical evaluation of software-as a medical device ([IMDRF, 2017a](#)) are not specifically aimed to AI-enabled software.

When conducting validation of AI systems, adequate consideration must be given to AI-specific aspects such as → data quality that is fit for purpose ([Griesinger et al., 2022](#)), → bias, modelling decisions (that may carry over biases or introduce algorithmic biases; → algorithm-to-model transition) and → intelligibility of AI systems.

### **Context dependency of validation**

Validation of AI-based systems is context dependent and needs to be tailored to the intended purpose: there are numerous possible → use environments and → use contexts in the health sector. Thus, a general “one-size-fits-all” type approach is not useful. As with all scientific studies the objectives of a study (including importantly the intended purpose) will determine the required design.

- Common challenges of validation include the difficulty of setting required level of assurance linked to establishing “validity” for the → intended use. While → model performance is obviously of key importance, AI system validation has a much broader scope than → model validation (typically of a machine learning model).
- AI system validation needs to address also issues related to the composite nature of the system, e.g. interplay of → AI system components / parts for the functioning of the system, performance of the system as a whole under controlled conditions.
- In contrast, aspects of usability or → clinical performance and → clinical effectiveness are examined during → clinical validation which should incorporate → usability validation. According to a recent literature review by Myllyaho et al (2021), the most frequently used validation methods of AI systems are 1) expert opinion, 2) simulation, 3) model-centred, 4) trial. The epistemological value increases in this order.
- The British standard BS30440 provides a validation framework for the use of AI in healthcare (summarised in Sujan et al., 2023), covering relevant topics including security vulnerabilities (→ cybersecurity cluster of terms) or explainability.

### **Challenges posed by AI-based systems**

A major source of challenges concerns the sheer scale and diversity of the intended → use environments and → use context for AI.

AI systems that continuously adapt their model based on → post-deployment input data (→ continuous and adaptive learning) pose a major challenge. While consistency of outputs/behaviour is usually viewed as a desired quality in software or AI systems under specific use conditions (e.g. diagnostic support), continuously learning and hence continuously adapting (and improving) output characteristics may be a desirable property (e.g. in health research). In such cases, setting target values or parameters of desired outcomes for purposes of validation will be challenging and other avenues of ensuring that the model performs adequately may be required (e.g. through → post-deployment monitoring).

Drifts and shifts (→ drifts / shifts), such as → data drift / shift, and → distributional drift / shift may complicate both, validation and continuous → post-deployment monitoring.

### **Term relationship:**

Related terms:

- Verification
- Model validation
- Usability validation
- Clinical validation

## **Usability validation**

**Cluster:** B.10 Verification, validation, evaluation

**Concept description**

**Usability validation** is a term that has been used in various technical areas including software and medical devices and is hence also applicable to AI systems in health and medicine (e.g. → **AI-enabled medical device software**).

In general terms, usability validation refers to the confirmation through objective evidence during use conditions that the AI system meets the user/usability requirements for the intended use. In case of → **AI-enabled medical device software**, usability validation and may be addressed in the context of → **clinical validation** ([WHO, 2021b](#)). For other AI systems (e.g. for health research, health systems planning, public health) that do not require a clinical validation, a dedicated usability validation may be recommended.

Depending on needs, usability validation

- sheds light on the facility or ease of use and thus provides information on → **user competency and training requirements**. This entails whether the user interface(s), required prompts or protocols are clear and allow efficient use of the device. User satisfaction should be part of usability metrics.
- informs on → **efficiency** of the AI system within the intended → **use context** and → **use environment**.
- Allows assessing whether usability is sufficiently satisfactory so as not to interfere with safety: poor usability may lead to user errors and possibly harm (→ **AI safety**).

In the context of usability validation, 'usability testing' refers to the actual methods for testing or evaluating usability with end users within a specified intended → **use environment**. For AI systems used for medical purposes (see also → **AI-enabled medical device software**), usability tests may be preclinical or clinical depending on the test design and validation needs (see also → **clinical evaluation**).

#### Explanatory note

##### **Usability and user interface: EU jurisdictions**

Notably, in the EU jurisdiction, medical devices usability validation mainly refers to the characteristic of the *user interface* that establishes effectiveness, efficiency and ease of user learning and user satisfaction ([MDCG 2020a](#)).

##### **Usability can impact safety**

Usability has repercussions concerning safety (→ **AI safety**): errors due to poor usability when using an → **AI-enabled medical device software** can lead to harm. Regulatory authorities are addressing the issue by implementing guidelines to improve usability engineering practices and minimize usability issues (see also [MDCG, 2020a](#); [MDCG, 2024a, b](#)).

##### **Usability engineering**

Manufacturers might want to integrate in their design and development lifecycle a safety-oriented usability engineering process ([IEC, 2015](#); [US FDA, 2016](#); [ISO, 2015b](#); [ISO, 2019](#)). A critical step in this process concerns identification of potential *use errors* (see also → **foreseeable misuse**) ([Schiro et al., 2017](#)). Use errors that have been reported for similar devices should be identified in order to avoid repeating foreseeable problems for the product under development.

#### Term relationship:

Related terms

- Usability
- Verification
- AI system validation ('analytical / technical validation')
- Clinical validation
- Usability validation
- User research

## Clinical validation

### Cluster: B.10 Verification, validation, evaluation

#### Concept description

Clinical validation aims to confirm that the medical device (e.g. → AI-enabled medical device software) is performant, is safe and has clinical utility. Evidence for clinical validation is generated in the context of dedicated clinical studies, → **clinical investigations** or clinical trials, ideally using prospective data.

For AI-enabled tools, clinical validation should be performed on the background of preliminary evidence from → **model validation**, → **model testing** and → **model evaluation**. Evidence generated in the context of clinical validation feeds into → **clinical evaluation**, covering both pre-and post-market space.

Studies aiming at clinical validation address the following topics:

- **Efficacy** (→ efficacy, effectiveness and efficiency), i.e. whether the device performs its intended clinical function under **controlled, optimal conditions**, e.g. whether an AI-enabled medical device achieves its diagnostic or therapeutic objectives in a research setting and shows → **clinical benefits**.
- **Safety**: whether the device is safe for patients, users and other persons (→ AI safety) under optimised conditions (e.g. research settings) and, more importantly, under real-world conditions (→ **real world use**).
- **Clinical performance** (→ clinical performance): does the device perform as intended (→ intended use) and does it provide the envisaged → **clinical benefits** to patients **under real world conditions**? This relates to broader concept of **effectiveness** (→ clinical effectiveness) capturing question such as: does the device perform satisfactorily with typical users (→ **use context**), diverse patient populations and in diverse → **use environments**? This entails dedicated studies involving different healthcare settings, clinicians with varying competency and skill levels.
- **Usability** (→ usability validation). This includes how easy it is for clinicians and healthcare professionals to use the device in actual clinical workflows. Usability may include how easy AI system outputs (→ **output and output data**) can be utilised by clinicians (e.g. to support diagnostic decision making) and how intelligible outputs are (→ **intelligibility**; → **causability**).

In summary, the systematic and rigorous validation of AI-enabled medical devices is an essential step before their integration into clinical workflows or even more standardised clinical practice or 'models of practice' (→ **clinical practice guideline**; → **clinical practice protocol**; → **clinical pathway**). Clinical validation ensures that the device performs as intended, is safe, and provides actual benefits to patients, while being usable by and of utility to the end users. Clinical validation can prevent unintentional harm, including misdiagnosis, inappropriate treatment or adverse effects with potential effects on patient well-being or even survival.

#### Explanatory note

##### **Clinical validation and clinical evaluation**

Note that the concept of clinical validation is included in the chapter on "evaluation of digital interventions / AI-SaMD" in WHO's guidance document ([WHO, 2021b](#)). This reflects the conceptual relationship between clinical validation (mostly through prospective data) and → **clinical evaluation** which includes a range of data and → **clinical evidence** collected during both the pre- and postmarket space, including retrospective data, postmarket surveillance data (→ **postmarket surveillance**, **market surveillance**, **corrective action**), post-market → **clinical evidence** (→ **post-market clinical follow-up**), user feedback and, importantly, prospective data from clinical validation, obtained in dedicated studies (e.g. clinical studies, clinical trials, → **clinical investigations**).

##### **Relevant guidance and publications on medical devices**

For guidance on medical devices that is relevant in this context, see [IMDRF, 2019a](#); [IMDRF, 2017a](#); [GHTF, 2012](#); [EU MDCG, 2019/2020a](#); [US FDA/Health Canada/UK MHRA, 2021](#); [WHO 2021b](#); [IMDRF 2025a, b](#); [EU MDCG, 2025](#). The paper by [Peabody et al. \(2014\)](#) provides useful reflections on clinical validation for molecular diagnostics, that of [Park et al. \(2021\)](#) on validation of AI systems.

#### **Difference to technical and analytical validation**

Clinical validation focuses on real-world clinical performance and utility, patient benefit and safety. In contrast, technical or analytical validation exercises (→ [AI system validation \('analytical / technical validation'\)](#)) examine predominantly technical performance aspects, e.g. accuracy for classification outcomes and should address the overall technical functioning of AI systems that draw on different components (→ [AI system component / part](#)), e.g. an AI-enabled surgical suite with sensors, actuators.

#### **Clinical validation and clinical investigations (or 'clinical trials', or 'clinical studies')**

Well-designed → [clinical investigations](#) in which AI-enabled medical devices are compared to the current gold standard (i.e. routine care delivered by medical experts) are a possible approach for assessing performance and safety. These studies provide a more detailed evaluation, including a range of relevant parameters, such as patient benefits in terms of quality of life, acceptance by physicians, integration into the clinical workflows, and economic impact. However, such studies are costly, both financially and in terms of time required and may be challenging to organise from a patient recruitment point of view (consider an AI-enabled diagnostic device for a rare disease). They should be preceded by early-phase studies. Further, by using current routine care as a gold standard or benchmark, there is a certain risk that models are aligned with imperfect practices involving human actors.

[Tsopra et al \(2021\)](#) has proposed a pragmatic approach to clinical validation for AI technologies: these should undergo clinical validation with external real-world datasets once they have achieved a state of sufficient "maturity", i.e. rather early during their life cycle. If the performance of the specific AI technology falls short of expectations at this early validation (e.g. if it fails to predict response to treatment, or is considered unsafe), it can be rejected (as in early-phase trials for drugs), and no further evaluation in RCTs is required. However, if an AI is successfully validated from a clinical perspective with these real-world datasets, it can be considered a good candidate and allowed to progress to the next stage in evaluation (i.e. a randomised controlled trial, RCT). This proposal appears to further extend the concept of → [model evaluation](#).

#### **Term relationship:**

Related terms:

- [Clinical evaluation](#)
- [Clinical investigation](#)
- [Efficacy, effectiveness, efficiency](#)
- [Model evaluation](#)
- [BENEFICENCE](#)
- [NON-MALEFICENCE](#) (risk avoidance and management)
- [Intended use](#)
- [Usability](#)

## Clinical evaluation

**Cluster:** [B.10 Verification, validation, evaluation](#)

#### **Concept description**

The term clinical evaluation is well-established in the area of medical devices and is in the present context relevant in particular for → [AI-enabled medical device software](#). This includes the concepts of a) AI-enabled software in a medical device (AI-SIMD, e.g. to control, drive, influence or support a medical device including for analysing data) and b) AI-enabled software as a medical device (AI-SAMD, e.g. an AI

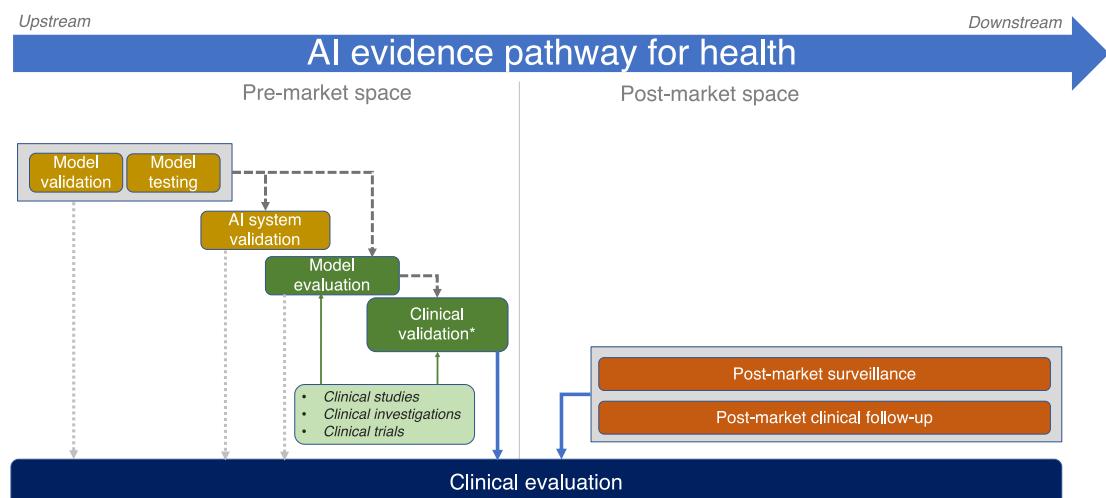
model supporting clinical decision-making by integrating relevant patient data and producing recommendations) (see [IMDRF, 2017](#)).

According to the IMDRF ([2019a](#)):

- “*Clinical evaluation is a set of ongoing activities that use scientifically sound methods for the assessment and analysis of clinical data to verify the safety, clinical performance and/or effectiveness of the device when used as intended by the manufacturer.*”
- “*Clinical evaluation is an ongoing process conducted throughout the life cycle of a medical device. It is first performed during the development of a medical device in order to identify data that need to be generated for regulatory purposes and will inform if a new device clinical investigation is necessary, together with the outcomes which need to be studied. It is then repeated periodically as new safety, clinical performance and/or effectiveness information about the medical device is obtained during its use. This information is fed into the ongoing risk management process (according to ISO 14971:2007) and may result in changes to the manufacturer's risk assessment, clinical investigation documents, Instructions for Use and post market activities.*”

**Figure 26.** Highly schematic representation of key validation and evaluation activities feeding into clinical evaluation of AI systems used in healthcare and medicine. Dark grey stippled arrows: validation activities that may inform subsequent evaluation and validation. Light grey stippled arrows: information may contribute to clinical evaluation. Green arrows indicate evidence generated during dedicated prospective studies in the context of model evaluation and clinical validation. Blue arrows indicate main contributions from clinical validation and relevant post-market information, e.g. post-market surveillance, post-market clinical follow up. Yellow boxes = technical validation/testing. Green boxes: evaluation and validation generating clinically relevant evidence. Orange boxes: information from the post-market space. Note that various impact assessments that may be required are not shown

### AI systems used in healthcare and medicine



\*) Clinical validation incorporates typically usability aspects (usability validation).

Source: own production

### Explanatory note

According to IMDRF ([2019](#)) the process of clinical evaluation is as follows:

“To conduct a clinical evaluation, a manufacturer needs to: • identify the Essential Principles that require support from relevant clinical data; • identify available clinical data relevant to the medical device and

*its intended use; • evaluate (appraise and analyses) clinical data in terms of its suitability and contribution to demonstrating the safety, clinical performance and/or effectiveness of the medical device in relation to its intended use; • generate clinical data needed to address remaining questions of safety, clinical performance and/or effectiveness; • bring all the clinical data together to reach conclusions about the safety, clinical performance and/or effectiveness of the medical device. The results of this process are documented in a clinical evaluation report. The clinical evaluation report and the clinical data on which it is based serve as the clinical evidence that supports the marketing of the device. The clinical evidence, along with other design verification and validation documentation, device description, labelling, risk analysis and manufacturing information, is needed to allow a manufacturer to demonstrate conformity with the Essential Principles and is part of the technical documentation of a medical device.”*

#### Term relationship:

Related terms:

- Clinical studies
- Clinical trial
- Clinical investigation
- Clinical evidence
- Post-market surveillance, market surveillance, corrective action
- Post-market clinical follow-up
- Continuous evaluation of AI systems

## Continuous evaluation of AI systems

#### Cluster: B.10 Verification, validation, evaluation

#### Concept description

This term concerns AI systems that are not used for healthcare, medical or clinical purposes but, instead, for other applications, e.g. health research, health system administration and planning or public health (categorisation according to [WHO, 2021a](#)). We propose this term as a ‘relative’ of → **clinical evaluation** for AI systems used in healthcare, medicine and clinical purposes.

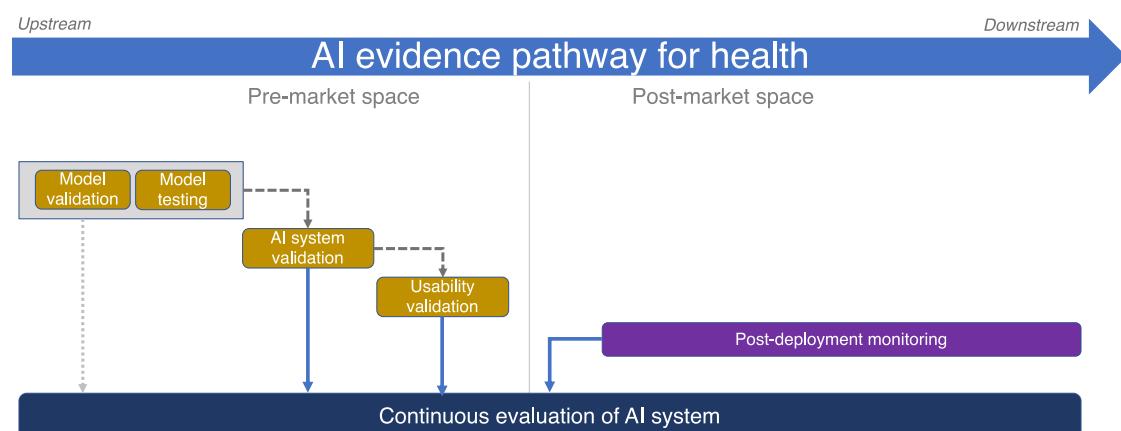
With ‘continuous evaluation of AI systems’, we refer to the ongoing collection and evaluation of information on AI system characteristics throughout their lifecycle and covering both the pre-and post-market space. Characteristics include

- Performance of model (→ **model performance**) and AI system
- Usability as examined during pre-market → **usability validation** and based on user feedback, once the AI system has been deployed.
- Any other relevant information obtained during → **post-deployment monitoring**.

Various validation and evaluation exercises should feed into the continuous evaluation of AI systems. This includes before deployment: → **model validation**, → **AI system validation** ('analytical / technical validation') and → **usability validation** and after deployment → **post-deployment monitoring** and → **peer review and community discourse**.

**Figure 27.** Highly schematic representation of key validation and assessment activities that, depending on AI system, may need to be considered for the continuous evaluation of AI systems that are not used in healthcare and medicine but for other applications such as health research, health system management and public health. Dark grey stippled arrows: validation and testing activities that may inform subsequent AI system validation and usability validation. Light grey stippled arrows: information may contribute to continuous evaluation of the AI system. Blue arrows indicate contributions of validation exercises and post-market information to the continuous evaluation of the AI system. Yellow boxes = technical validation/testing. Purple box: information from the post-market space. Note that various impact assessments that may be required are not shown.

AI systems used for health research, health system management, public health



Source: own production

#### Term relationship:

Related terms:

- Clinical evaluation
- Model validation
- AI system validation ('analytical / technical validation')
- Usability validation
- Post-deployment monitoring
- Peer review and community discourse

## B.11 Clinical concepts

### Health outcomes

**Cluster:** B.11 Clinical concepts

#### Concept description

Health, according to the WHO's constitution ([1946; several amendments](#)) is "*a state of complete physical, mental, and social well-being, and not merely the absence of disease or infirmity*". Consequently, health outcomes can cover a wide variety of health-related consequences, typically resulting from interactions of individuals or populations with healthcare systems ([Lee & Leung, 2014](#)). This can include prevention, screening, early diagnosis and treatments ('cure') – the latter having been traditionally in the focus of health outcomes.

#### Explanatory note

- Due to demographic changes in rapidly aging societies (e.g. EU, China), *prevention* is rapidly moving into the focus of healthcare systems and healthcare providers ([Badimon et al., 2023](#)) as a preferred means of achieving positive health outcomes and avoiding therapeutic interventions at late stages that causes patient discomfort and come at great socioeconomic cost ([Healthcare-in-europe.com, 2024](#)).
- Lee & Leung (2014) have defined health outcomes as follows: "*Health outcomes refer to the health consequences brought about by the treatment of a health condition or as a result of an interaction with the healthcare system. It is a multidimensional concept that can be studied on multiple levels.*"

#### Term relationship:

Related terms:

- BENEFICENCE
- Clinical performance
- Clinical effectiveness
- Clinical benefit

### Healthcare modality

**Cluster:** B.11 Clinical concepts

#### Concept description

Clinical or healthcare modality refers to types of healthcare relevant actions, e.g. diagnosis or treatment, an interventional strategy or, more specifically, a piece of equipment. An example of modalities from oncology are chemotherapy, radiation therapy or surgery. AI tools can support various healthcare modalities, e.g. diagnostic and prognostic predictions, clinical decision making in relation to the optimal choice of treatments etc. Healthcare modalities should not be confused with → **data modality**, although there may be a certain overall. For example in radiology, medical imaging modalities refer to data from specific pieces of equipment (e.g. MRI, CT scans, ultrasound and x-rays).

## Efficacy, effectiveness and efficiency

### Cluster: B.11 Clinical concepts

#### Concept description

The terms **efficiency**, **efficacy** and **effectiveness** have slightly different notions in healthcare and AI in health ([Burches & Burches, 2020](#)):

- **Efficacy** relates to the effectiveness of health interventions as demonstrated in **research settings**, e.g. changed processes or output measures, changed outcomes (e.g. healthcare worker performance, patient health outcomes) ([WHO, 2021](#)). Efficacy is typically evaluated under controlled and optimised conditions of clinical studies.
- **Effectiveness** is used to capture the impact of AI-based devices on process change in the **intended setting** (e.g. time to diagnosis), outcomes (e.g. user accuracy/efficiency, patient health outcomes), and cost-effectiveness ([WHO, 2021](#)).
- **Efficiency** has a slightly broader scope than effectiveness. It can be considered one of the three ‘cardinal’ design features that a health product must fulfil (in the present context an → **AI-enabled medical devices software**) with the other two being safety and equity ([Char et al., 2021](#)). **Efficiency** captures the notions
  - whether and to which extent the AI system **addresses and solves the (clinical) problem** it was designed for, i.e. shows intended benefits.
  - whether it does this at a reasonable **economic cost**. Importantly, cost is not only a matter of pricing of the marketable product, but also of **collateral costs**, e.g. costs in new/updated **infrastructure**, secondary costs that may arise due to **cybersecurity vulnerabilities**, cost incurred due to **training needs** and/or **poor usability** etc.
  - with **acceptable error rates** (e.g. false negatives (FN), false positives (FP)). What is considered ‘acceptable’ depends very much on the intended use and integration of the system in a given workflow. For instance, in case the system is intended for cancer screening, a comparably high rate of FN would be problematic while a higher rate of FP might be acceptable).

#### Explanatory note

Aspects of efficacy and effectiveness are captured in this ontology under the ethical principle of → **BENEFICENCE**. We make a distinction between **potential gains** (aimed at in the pre-market development phase), **potential benefits** (in research settings in the pre-market space) and **real-world benefits** (in the post-deployment phase / intended setting).

#### Term relationship:

Related terms:

- **BENEFICENCE**
- **Clinical performance**
- **Clinical effectiveness**
- **Clinical benefit**

## Clinical effectiveness

### Cluster: B.11 Clinical concepts

#### Concept description

Clinical effectiveness is a key concept of evidence-based medicine (EBM). It concerns the 'effectiveness' of a specific treatment or intervention, i.e. to allow to answer the question whether a given intervention

- actually works and, in addition,
- whether it works better than another one, i.e. comparative aspect ([Ashcroft, 2002](#)).

The term "works" in this context can have several meanings, including from a purely clinical perspective (e.g. does the treatment decrease blood pressure?) but also from a perspective of patient-reported well-being (do the patients feel better?).

Like → [clinical performance](#), clinical effectiveness is a relational term, i.e. effectiveness of a treatment can be shown in a meaningful way only in relation to an → [intended purpose](#) involving also possibly a specific → [use context](#) and → [use environment](#).

Clinical effectiveness can relate to three domains: a) clinical investigations (e.g. of a → [AI-enabled medical device software](#)) or randomised clinical trials (RCT) in case of medicinal products; b) clinical practice under real-world conditions (as opposed to the optimised conditions of an investigation or RCT; see also → [efficacy, effectiveness and efficiency](#)) and c) wider health policy.

Information from → [post-deployment monitoring](#) and → [post-market surveillance](#) may provide additional information to clinical effectiveness and add clinical and patient-based experiences obtained under real-world conditions. Clinical effectiveness relates to the concept of → [clinical benefit](#) and thus to the ethical principle of → [beneficence](#). Evaluations and measurement of clinical effectiveness are relevant also for health technology assessment (→ [health technology assessment](#)), in particular the comparative aspect of clinical effectiveness.

#### Explanatory note

IMDRF defines clinical effectiveness as: "*The ability of a medical device to achieve clinically meaningful outcome(s) in its intended use as claimed by the manufacturer*" ([IMDRF, 2019a, 2019b](#)).

#### Term relationship:

Related terms:

- Intended use
- Clinical investigation
- Clinical evaluation
- Clinical performance
- Clinical outcome

## Clinical performance

### Cluster: B.11 Clinical concepts

#### Concept description

In general, clinical performance refers to the extent to which a health product (e.g. medical device, → [AI-enabled medical devices software](#)) achieves its → [intended use](#) ([IMDRF, 2019a, b, c](#)) and provides → [clinical benefits](#) to patients under real-world conditions. Thus, clinical performance typically entails considerations of → [use context](#) and → [use environment](#).

Thus, clinical performance is a relational term and cannot be obtained or measured for the product on its own. Clinical performance may be influenced by → usability. Poor usability may negatively affect clinical performance.

IMDRF (2019a, document N56) defines clinical performance as “*The ability of a medical device to achieve its intended clinical purpose as claimed by the manufacturer.*”

#### Explanatory note

Clinical performance and safety of a device (e.g. → AI-enabled medical devices software) are addressed during → clinical evaluation and → clinical investigations. Additional information on → clinical effectiveness may add significant value in this context.

#### Term relationship:

Related terms:

- Intended use
- Clinical investigation
- Clinical evaluation
- Clinical effectiveness
- Clinical outcome

## Clinical benefit

### Cluster: B.11 Clinical concepts

#### Concept description

The benefit for a patient resulting from the use of a medical device. Clinical benefit should be described in terms of clinically relevant and measurable parameters. In the present context, clinical benefit is relevant for → **AI-enabled medical devices software**.

The EU MDR (Article 2, paragraph 53) defines clinical benefit as “*The positive impact of a device on the health of an individual, expressed in terms of a meaningful, measurable, patient-relevant clinical outcome(s), including outcome(s) related to diagnosis, or a positive impact on patient management or public health.*” (EU, 2017a).

#### Explanatory note

Clinical benefits are important for planning and conducting clinical investigations: these must be designed in such a way that potential remaining risks to subjects or third persons, after risk minimization, are justified when weighed against the clinical benefits to be expected.

#### Term relationship:

Related terms:

- Clinical risk
- Benefit risk ratio

## Clinical safety

### Cluster: B.11 Clinical concepts

#### Concept description

Clinical safety concerns primarily the safety of patients (→ patient safety), healthcare professionals and other persons. Clinical safety requires to avoid harm to persons that may result from procedures, processes or products used in a clinical or healthcare context. Notably, procedures, processes and products may have residual acceptable risks (see → AI safety). Clinical safety is achieved by anticipating, controlling and mitigating *avoidable* clinical risks (as opposed to legal or financial risks) through appropriate risk management procedures (→ AI risk management).

In the context of AI systems used in healthcare, we cover clinical safety under the term → AI safety, in order to emphasize the specific safety challenges of AI.

#### Explanatory note

The IMDRF guidance document on “essential principles of safety and performance of medical devices and IVD medical devices” (IMDRF, 2018, second edition 2024) refers to clinical safety. The document does not specifically address *AI-enabled* medical and IVD medical devices.

#### Term relationship:

Related terms:

- AI safety
- Patient safety
- AI risk management

## Patient safety

### Cluster: B.11 Clinical concepts

#### Concept description

Patient safety is a key aspect of → clinical safety. Patient safety has been defined by WHO as “*the absence of preventable harm to a patient and reduction of risk of unnecessary harm associated with health care to an acceptable minimum.*” (WHO, 2009).

In the context of AI systems used in healthcare, we discuss patient safety under the term → AI safety, in order to emphasize the specific safety challenges of AI.

#### Explanatory note

Apart from being a desirable state for patients, patient safety is also a practical discipline in healthcare aiming to prevent, reduce, report and analyse errors or other sources of unnecessary harm or risks (Emanuel L et al., 2008).

#### Term relationship:

Related terms:

- AI safety
- Patient safety
- AI risk management

## Clinical evidence

### Cluster: B.11 Clinical concepts

#### Concept description

Clinical evidence is a term used across various medical and health domains, including medical devices, medicinal products (drugs), diagnostic tests, specific procedures (e.g. surgery, rehabilitation, physiotherapy) and treatment protocols.

Clinical evidence can be derived by critical analysis of a broad spectrum of information sources, including, most importantly, → clinical investigations (or clinical trials), patient outcomes (including patient-reported outcomes), systematic reviews and meta-analyses of the medical literature, data on safety and efficiency, including from → post-market clinical-follow-up (PMCF), long-term follow-up studies and post-market surveillance activities (→ post-market surveillance, market surveillance, corrective action).

In the context of AI tools in healthcare and, specifically, → AI-enabled medical device software, the collection of relevant data and their potential qualification as robust clinical evidence plays a key role for assessing and establishing → clinical safety, → patient safety, → AI safety, → clinical benefit, → clinical performance and → clinical effectiveness. Clinical evidence concerns both, the → clinical data generated and the analysis of that clinical data in the context of the continuous process of → clinical evaluation.

For details see: [IMDRF 2012, 2017, 2019 and EU MDCG, 2020](#).

#### Explanatory note

According to IMDRF (N56, 2019) clinical evidence “*along with other design verification and validation documentation, device description, labelling, risk analysis and manufacturing information, is needed to allow a manufacturer to demonstrate conformity with the Essential Principles and is part of the technical documentation of a medical device.*”

#### Term relationship:

Related terms:

- Clinical data
- Clinical evaluation

## Clinical data

### Cluster: B.11 Clinical concepts

#### Concept description

According to [IMDRF \(2019: N55, N56, N57\)](#) clinical data relates to information on the safety, → clinical performance, and/or → clinical effectiveness generated from the clinical use of a medical device; in the present context this concerns → AI systems used in healthcare, i.e. → AI-enabled medical device software.

#### Explanatory note

Possible sources of clinical data include:

- (i) pre- and post-market clinical investigation(s) of the concerned AI system / device
- (ii) pre- and post-market clinical investigation(s) or other studies reported in the scientific literature of a sufficiently comparable device
- (iii) published and/or unpublished reports on clinical experience of either the concerned AI system / device or a sufficiently comparable device

|   |
|---|
| (iv) registries, adverse event databases or medical records |
| <b>Term relationship:</b>                                   |

Related terms:

- Clinical evidence
- Clinical performance
- Clinical effectiveness
- Clinical investigation
- Post-deployment monitoring
- Post-market surveillance, market surveillance, corrective action

## Clinical investigation

### Cluster: B.11 Clinical concepts

#### Concept description

According to [IMDRF \(2019a\)](#), a clinical investigation is a systematic investigation or study involving human subjects with the aim of investigating the safety, clinical performance and / or effectiveness of a medical device and, in the present context, of a → AI-enabled medical device software.

Clinical investigations may also be conducted in the context of → model evaluation or various activities aimed at the evaluation of digital interventions / AI-SaMD ([WHO, 2021b](#)). Clinical investigations are important for generating sufficient evidence and evidence of sufficient quality for → clinical validation and → clinical evaluation ([IMDRF, 2017; 2019b; EU MDCG, 2020](#)).

#### Explanatory note

Clinical investigations for medical devices and clinical trials for medicinal products are necessary to provide data not available through other sources (such as literature or non-clinical testing) required to demonstrate compliance with the relevant regulatory framework applicable in each jurisdiction. For example for AI-SaMD systems the relevant medical devices regulatory framework will apply in conjunction with any other applicable legislation e.g. the AI-Act in the European jurisdiction. In this case the data obtained will be used in the clinical evaluation process and will be part of the clinical evidence for the AI-SaMD.

In any case clinical investigations must consider the scientific principles underlying the collection of clinical data along with accepted ethical standards surrounding the use of human subjects. For example, ISO 14155: 2020 "Clinical Investigation of Medical Devices for Human Subjects - Good clinical practice" ([ISO, 2020](#)) details the requirements for the conduct of clinical investigations in the area of medical devices being. This standard is widely considered the equivalent to the ICH CGP E6 guideline for the proper design, conduct and reporting of clinical trials with medicinal products.

#### Term relationship:

Synonyms:

- Clinical trial\*
- Clinical study\*

Related terms:

- Clinical evaluation

*\*) N.B. Both IMDRF and EU legislative texts refer exclusively to „clinical investigations“ as dedicated studies of the clinical benefits, risks and performance of a medical device. However, in practice, the*

terms „clinical trial“ and „clinical study“ are used interchangeably with clinical investigation in various documents (see for instance ISO 14155:2011).

## Clinical endpoint

### Cluster: B.11 Clinical concepts

#### Concept description

[IMDRF \(2019c\)](#) defines clinical endpoint as: *“An indicator used for providing the evidence for safety, clinical performance, and/or effectiveness in a clinical investigation.”*

#### Explanatory note

Clinical endpoints are considered as principal indicators used for assessing the primary or secondary hypothesis of a clinical investigation. According to EU's MDR (point 2.6 in Annex XV on clinical investigations; [EU, 2017a](#)): *“The endpoints of clinical investigations shall address the intended purpose, clinical benefits, performance and safety of the device. The endpoints shall be determined and assessed using scientifically valid methodologies. The primary endpoint shall be appropriate to the device and clinically relevant.”* ([EU, 2017](#)).

#### Term relationship:

Related terms:

- [Clinical investigation](#)
- [Clinical trial](#)

Synonyms:

- [Clinical outcome](#)

## Clinical investigation plan

### Cluster: B.11 Clinical concepts

#### Concept description

A clinical investigation plan (CIP) is an important part of a medical device → [clinical investigation](#) (or clinical trial). The CIP outlines all elements that need to be considered for conducting the clinical investigation, including the sponsor and funding, design of the clinical investigation, e.g. number of subjects, inclusion/exclusion criteria, duration, follow-up, clinical endpoints (parameters recorded), the overall rationale and objectives.

#### Explanatory note

The EU medical devices coordination group has published guidance on the context of the clinical investigation plan for clinical investigations of medical devices ([EU MDCG, 2024](#)). Other available documents on clinical investigation are by IMDRF (2019) and ISO ([ISO standard 5, 2020](#)).

Relevant guidance on key elements for a CIP for medical devices can be found in IMDRF guidance N57 (2019). ISO has published a standard (14155:2020) on clinical investigations of medical devices for human subjects (ISO, 2020). Note should be taken that available guidance is aimed at devices in general. Specific aspects relating to AI-enabled medical purpose software are not outlined.

Tsopra et al. (2021) have published a paper on clinical validation of AI in precision medicine, which includes additional elements that are specific to AI, e.g. datasets used, data safety (including quality, privacy, security), AI performance measurement and, importantly, AI explainability.

### Term relationship:

Related terms: Text

- Clinical evaluation
- Clinical investigation
- Clinical study / trial

## Clinical predictions

### Cluster: B.11 Clinical concepts

#### Concept description

Clinical predictions is an umbrella term for predictions made in healthcare contexts and clinical practice. Clinical predictions are often formalised as clinical prediction rules. These aim at reducing the uncertainty of medical practice by defining how to use clinical observations and findings in a consistent and correct manner (Wasson et al., 1985).

Clinical predictions cover a wide range of patient-related forecasts based on a broad array of available medical information. Every year, new clinical prediction models are being published. However, many of these may be limited in their usefulness due to lack of methodological standards (Cowley et al., 2019) when developing and evaluating these, e.g. a lack of external validation (Efthimiou et al., 2024).

Clinical prediction models may support physicians in deciding whether a specific diagnostic test is required (a classic example is the “Ottawa ankle rules”; Mayer, 2009), help with the classification of patients for treatment (e.g. Beattie & Nelson, 2006), facilitate treatment decisions (e.g. Giles-Clark et al., 2023), support risk assessments and predictions of treatment response.

Clinical predictions include both *diagnostic predictions* (determination of the probability of a specific disease/condition in a patient, classification of disease type or stage, differential diagnosis) and *prognostic predictions* (estimation of future health outcomes, of disease progression (including morbidity), of complications, of survival rates) (van Smeden et al., 2021).

#### Explanatory note

AI-enabled tools may be of huge benefit for developing and evaluating clinical prediction models, due to capacity to analyse large amounts of data swiftly, to detect patterns that might go unnoticed by human readers and by their *potential* objectivity, i.e. lack of heuristic decision-making patterns (→ **heuristics**) – this objectivity depends however on many factors that need to be conscientiously chosen (→ **algorithm-to-model transition**). Examples for AI-enabled clinical prediction tools: Tarabanis et al., 2023 (atrial fibrillation), Pezel et al., 2023 (heart failure).

When developing clinical prediction models relevant available resources such as the PROGRESS framework ([prognosisresearch.com](https://www.prognosisresearch.com), 2025) and the AI versions of the PROBAST and TRIPOD frameworks (see Annex 1) should be used (Efthimiou et al., 2024).

## Term relationship:

Related terms: Text

- Diagnostic prediction
- Prognostic prediction
- Clinical decision-support tools
- **Clinical practice guideline (CPG)**
- **Clinical pathway**

## Adverse event

### Cluster: B.11 Clinical concepts

#### Concept description

Generally, adverse event means any untoward medical occurrence in patients/subjects, users or other persons resulting from a medical treatment/intervention. In the present context this encompasses events resulting from the use of a medical device that is AI-enabled.

In several jurisdictions, the term adverse event covers both the post-market space as well as clinical investigations (pre-market space) (see [Global harmonisation task force, now IMDRF, 2006](#)). However, in the EU, the term adverse event is restricted to clinical investigations, while adverse events during the post market phase are called → [incidents \(EU, 2017a and 2017b\)](#). Thus, depending on jurisdiction, the term adverse event (in its post-market meaning) and incident can be used interchangeably.

#### Explanatory note

In the context of clinical investigations, for patients/subjects, adverse events includes all untoward medical occurrences that occurred in the course of the investigation, whether or not related to the device under clinical investigation ("investigational device"). In contrast, in the context of clinical experience (outside a clinical investigation), adverse event includes untoward medical occurrences that may be related to the medical device. The possible association between device use and adverse event typically needs to be carefully investigated (see root cause terminology subset of the IMDRF's adverse event terminology: [IMDRF, 2020a](#)).

## Term relationship:

Related terms:

- **Incident**

Synonyms (depending on jurisdiction)

- **Incident**

## Incident

### Cluster: B.11 Clinical concepts

#### Concept description

In the European jurisdiction incident is defined in the EU's medical devices Regulation ([EU, 2017](#)) as "*any malfunction or deterioration in the characteristics or performance of a device made available on*

*the market, including use-error due to ergonomic features, as well as any inadequacy in the information supplied by the manufacturer and any undesirable side-effect”.*

#### Explanatory note

- Note should be taken that the term *adverse event* is used in some jurisdictions for events during clinical trials as well as in the post-market space, whereas in the EU the term ‘incident’ is used for adverse events that happen in the post-market space, with the term ‘adverse event’ used predominantly to denote events during clinical investigations (pre-market space). Thus, “*depending on jurisdictions, the term adverse event (in its post-market meaning) and incident can typically be used interchangeably.*” ([IMDRF, 2020a](#)).
- Internationally agreed adverse event terminologies have been developed by the International Medical Devices Regulators Forum ([IMDRF, 2020a](#)) for reporting of adverse events and, in the EU, ‘incidents’ (see ‘manufacturers incident reporting form’ ([EU, 2020: so-called MIR form](#)))
- A web browser of the IMDRF terminologies has been developed by the EU Commission’s Joint Research Centre and handed over to IMDRF ([IMDRF, 2020b](#)).

Notably, the term “adverse event” in the context of clinical trials (i.e. in the pre-market space) has a more restricted meaning (c.f. GHTF/SG5/N5:2012) than in the post-market space (see above and GHTF/SG2/ N54R8:2006).

#### Term relationship:

Related terms:

- **Adverse event**

Synonyms:

- **Adverse event** (depending on jurisdiction)

## Post-market surveillance, market surveillance, corrective action

#### Cluster: B.11 Clinical concepts

#### Concept description

Relevant studies conducted in the pre-market space (e.g. clinical investigations, see [IMDRF, 2019](#)) should ensure that medical devices (including → AI-enabled medical device software) are safe and perform as intended when put on the market.

However, once placed on the market, the safety and proper performance of devices needs to be continuously monitored. Post-market surveillance and market surveillance relate to these monitoring activities in the post-market space. In the EU, post-market surveillance relates to activities of manufacturers and market surveillance to those of authorities. Both activities are important for ensuring the essential principles of safety and performance ([IMDRF, 2018](#); EU medical devices Regulation, Annex I: [EU, 2017](#)), once a product has been placed on the market. Post-market surveillance can be considered part of → **clinical evaluation**, which covers both the pre-market space and post-market space and is concerned with safety, performance and/or effectiveness.

**Post-market surveillance** typically encompasses the monitoring of:

- serious incidents and non-serious incidents (or, depending on jurisdiction, adverse events) and data on any undesirable side-effects
- trend reports (analyses of frequency of monitored events along a time-line) which support detection of patterns concerning undesired events and, possibly, root causes underlying these

(e.g. updating an algorithm of AI system relates to increased frequency of events; there may be delays between causes and monitorable effects)

- scientific or clinical literature, databases and/or registries/registers
- feedbacks and complaints by various actors (e.g. patients, healthcare professionals)
- available information about similar products.

Relevant requirements of reporting post-market surveillance information are outlined in relevant legislations (e.g. EU's medical devices Regulation, Chapter VII, section 2 'Vigilance'; [EU, 2017](#)). Guidance on tools such as terminology or templates for reporting of incidents/adverse events is available ([IMDRF, 2020a, 2020b](#); [EU Commission, 2020](#)).

Market surveillance encompasses various activities and measures to ensure that there are no health and safety issues or other aspects relating to the public interest.

#### Explanatory note

We refer here to the definitions provided by the EU' Regulation on medical devices ([EU, 2017](#)):

- **"Post-market surveillance"** means all activities carried out by manufacturers in cooperation with other economic operators to institute and keep up to date a systematic procedure to proactively collect and review experience gained from devices they place on the market, make available on the market or put into service for the purpose of identifying any need to immediately apply any necessary corrective or preventive actions.
- A post-market surveillance plan should be developed and be readily available before the public and wide-spread use of the AI system in order to monitor the **safety** and **clinical performance** of the AI system in a real-world setting and timely detect any issues that may arise after the deployment."
- **"Market surveillance "**means the activities carried out and measures taken by competent authorities to check and ensure that devices comply with the requirements set out in the relevant Union harmonisation legislation and do not endanger health, **safety** or any other aspect of public interest protection."
- **"Corrective action"** means action taken to eliminate the cause of a potential or actual non-conformity or other undesirable situation."

#### Term relationship:

Related terms:

- Adverse event
- Incident
- Clinical evaluation
- Post-deployment monitoring (for AI systems used outside a medical/healthcare context)

## Post-market clinical follow-up (PMCF)

#### Cluster: B.11 Clinical concepts

#### Concept description

Post-market clinical follow-up (PMCF) refers to the continuous collection and evaluation of clinically-relevant information of medical devices\* once these have been put on the market (i.e. in the 'post-market space') and are used in agreement with their → intended use in patients. PMCF feeds into the con-

tinuous process of → clinical evaluation and can be seen as part of post-market surveillance (PMS) activities (→ post-market surveillance, market surveillance, corrective actions). PMCF aims to gather information on whether

- safety and performance continue to be as found during the clinical evaluation before the device was put on the market
- the risks identified during pre-market evidence and studies of → clinical evaluation are still acceptable when analysing experiences under → real-world use conditions
- there are novel risks that were not identified before, including unknown effects and contraindications. Such indications should feed into updated post-market surveillance activities.
- there is misuse of off-label use of the device (which could have impacts on safety (see → AI safety))

*\*) in the present context an AI-enabled device or → AI-enabled medical device software*

#### Explanatory note

According to the EU MDR ([EU, 2017a](#)) “Post-market clinical follow-up shall be understood to be a continuous process that updates the clinical evaluation... and shall be addressed in the manufacturer’s post-market surveillance plan” ([EU, 2017a](#), Annex XIV, part B). Post-market clinical follow-up activities need to be conducted by the manufacturer of the device.

PMCF serves to update the continuous process of → clinical evaluation, with “*the aim of confirming the safety and performance throughout the expected life cycle of the device, of ensuring the continued acceptability of identified risks and of detecting emerging risks on the basis of factual evidence.*” ([EU, 2017a](#), Annex XIV, part B).

The findings of PMCF should be documented in a PMCF report, to be included in the → clinical evaluation report (CER) and the technical documentation ([EU, 2017a](#)).

PMCF can be also used as input to post-market surveillance (e.g. for updating the original post-market surveillance plan, PMS) and for updating not only the CER but also the summary of safety and clinical performance (SSCP).

#### Term relationship:

Related terms:

- Clinical evaluation
- AI safety
- Post-market surveillance, market surveillance, corrective action
- NON-MALEFICENCE (risk avoidance and management)

## Personalised medicine & precision medicine

### Cluster: B.11 Clinical concepts

#### Concept description

Personalized medicine (PM) seeks to overcome the traditional "one-size-fits-all" approach in medicine, by advancing individualised healthcare through tailored risk assessment, prevention and treatment strategies for *specific groups of individuals* or even for *single patients*. PM draws on the integration of diverse data including clinical history, lifestyle factors and a wide range of clinical data, notably genome data but also other omics data, imaging data, in vitro diagnostic data and other suitable biomarkers.

This may allow to design, either for groups or individuals, more targeted prevention measures and treatment choices, including on the basis of robust estimations of treatment efficacy and likelihood of

adverse events. Artificial intelligence solutions with their capacity of quickly analyzing large data sets are ideally suited to foster the concept of PM which aims to exploit the inherent data richness and multi-modality of health. This includes notably → generative AI models.

N.B. The terms 'personalised medicine' and 'precision medicine' are often used interchangeably in the literature, although, depending on authors, there may be subtle usage nuances (see explanatory note).

#### Explanatory note

- **Personalised medicine:** There is no universally accepted definition of personalised medicine available. The EU's Horizon 2020 Advisory Group defines personalised medicine as '*a medical model using characterisation of individuals' phenotypes and genotypes (e.g. molecular profiling, medical imaging, lifestyle data) for tailoring the right therapeutic strategy for the right person at the right time, and/or to determine the predisposition to disease and/or to deliver timely and targeted prevention.*' (see [European Commission, 2025c](#); [European Union, 2015](#)).
- **Precision medicine:** US National Human Genome Research Institute defines precision medicine as follows "*Precision medicine (generally considered analogous to personalized medicine or individualized medicine) is an innovative approach that uses information about an individual's genomic, environmental, and lifestyle information to guide decisions related to their medical management. The goal of precision medicine is to provide a more precise approach for the prevention, diagnosis, and treatment of disease.*" ([Delpierre & Lefèvre, 2023](#)).
- Notably, while above definition of personalised medicine refers to data from individuals (i.e. plural), above definition of precision medicine refers to information from "an individual" (i.e. singular), possibly suggesting a narrower focus of precision medicine on medical measures tailored to one individual only and based entirely on information from that individual.
- While, PM solutions are being integrated into healthcare, there remain substantial challenges, including from a regulatory and health technology assessment perspective, from an infrastructure perspective and a general policy perspective. There are also concerns that PM could sharply exacerbate existing health inequalities ([Brothers & Rothstein, 2015](#)). For an overview on PM see for [Vicente et al., \(2020\)](#) and [Rogers \(2025\)](#), for considerations of health policy and PM, see [Bureau et al., \(2021\)](#), for health reimbursement considerations, see [Koleva-Kolarova et al. \(2022\)](#). For a critique of PM that may promote a reductionist health model, see [Delpierre & Lefèvre, 2023](#).

#### Term relationship:

Related terms:

- BENEFICENCE, gains. In particular → "precision gain" and → "healthcare gain – personalisation of prevention/treatment"
- Health technology assessment
- Generative AI
- Foundation models

## B.12 Use of AI systems in health and healthcare

### Use

**Cluster:** B.12 Use of AI systems in health and healthcare

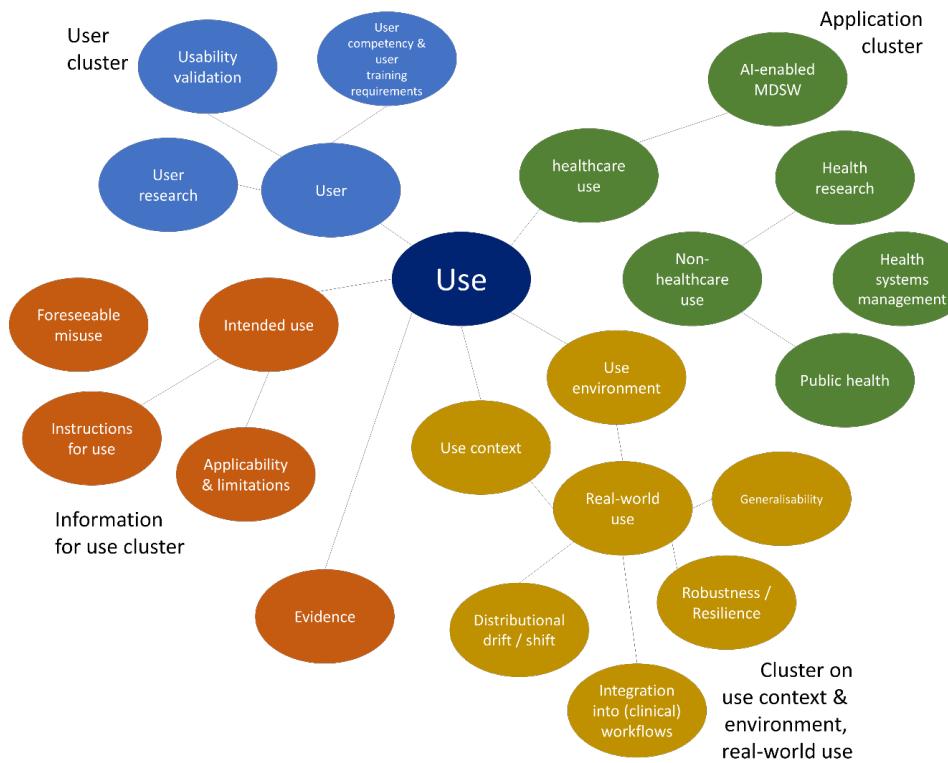
#### Concept description

AI systems in medicine and healthcare can be used in various ways. The WHO ([WHO, 2021](#)) has proposed a useful categorisation for applications of AI in healthcare:

- In **healthcare** (e.g. for diagnostics, clinical decision making, disease management, including through → **conversational agents**)
- In **health research and drug development**
- In **health systems management and planning** (e.g. use of → **conversational agents** for managing patient flows or emergency triaging; for management of electronic health records)
- In **public health and public health surveillance**

The terms use or application are connected to many other terms. Some of these can be found in this section ([Figure 28](#)).

**Figure 28.** Schematic depiction of the term 'use' and main related concepts. There are four clusters: 1) user, 2) application in health (as outlined in WHO, 2021). 3) Information for use, 4) Use context, use environment, real-world use.



Source: own production

#### Term relationship:

##### Synonyms

- Intended purpose

## User research

### Cluster: B.12 Use of AI systems in health and healthcare

#### Concept description

With user research we refer to a structured approach of investigating user needs during the concept and design phase of AI system so as to ensure that the final product meets indeed the user requirements (typically in a specific → use context or → use environment) and fulfils usability requirements (e.g. overall facility of use, user interface etc.) which may need to be confirmed subsequently through → usability validation.

User research may also help avoiding potential use errors due to so-called 'human factors' ('usability engineering', see ISO, 2015) to identify sources of misuse (→ foreseeable misuse) and ensure a positive user experience (see for instance: Borsci & David, 2020; Bitkina et al., 2020; Song et al., 2020).

For AI systems used in healthcare (e.g. → AI enabled medical device software), we understand user as defined in the EU's medical devices Regulation (EU, 2017a):

*"user" means any healthcare professional or lay person who uses a device" and*

*"lay person" means an individual who does not have formal education in a relevant field of healthcare or medical discipline"*

## Usability

### Cluster: B.12 Use of AI systems in health and healthcare

#### Concept description

Usability in the context of AI systems in health and medicine and in particular for healthcare (→ AI-enabled medical device software) refers to various aspects around the ease of use of an AI system for a specific → intended use and within a → use context (or "workflow") and → use environment. Results from → user research should inform usability considerations.

Importantly, usability is linked to → AI safety (see also → NON-MALEFICENCE), since poor usability can cause harm due to use errors or unnecessary complications or delays when using an AI system. Main concepts of usability include:

- **Efficiency and facility of use:** the AI system is not overly complex to use, e.g. the user interface is self-explanatory and has easy-to understand menu guidance). The usability supports → efficiency of the AI system in regard to the problem it was designed to solve. A → usability validation is required to generate evidence on efficiency and facility of use, including potential errors and unintended consequences.
- **User training requirements** are not disproportionate in regard to the expected benefits of running the AI system (e.g. no extensive barriers to reach proficiency of use; easy to understand → instructions for use; introductory training material and "on-boarding")
- **Ease of integration of the AI system into a specific workflow** or → use context
- **Ease of integration into a → use environment**, including flexibility with regard to differences of use environments concerning → IT infrastructure and → enabling technologies.

#### Explanatory note

The impacts of usability should also be considered in the context of (global) → health equity, i.e. to which extent is the usability supports a generalizable use for healthcare in a variety of settings and under a variety of circumstances that may deviate from the optimised conditions under which the system was developed and tested.

## Term relationship:

Related terms:

- Efficiency, efficacy, effectiveness
- Usability validation
- BENEFICENCE
- Clinical performance
- Clinical effectiveness
- Clinical benefit

## User competency and training requirements

### Cluster: B.12 Use of AI systems in health and healthcare

#### Concept description

With the concept 'user competency and requirements' we refer to the level of user competency required to operate the AI system safely and in order to ensure performance as foreseen. After determination of the required user competency, AI system developers and deployers should determine the level of training required to attain the level of user competency.

#### Explanatory note

User training can be provided via different routes, e.g. training manuals, training videos, training courses with proficient experts etc.

#### Term relationship:

Related terms:

- User research
- Usability
- Usability validation

## Intended use

### Cluster: B.12 Use of AI systems in health and healthcare

#### Concept description

The way how an AI system is supposed to be used.

Regarding AI-enabled medical device software, we refer to the IMDRF (IMDRF, 2019a) definition of "intended use" for medical devices and in vitro diagnostic medical devices: "*The objective intent regarding the use of a product, process or service as reflected in the specifications, instructions and information provided by the manufacturer.*"

#### Term relationship:

Related terms:

- Intended purpose

## Foreseeable misuse

### Cluster: B.12 Use of AI systems in health and healthcare

#### Concept description

Use of an AI system outside the proper use specifications as specified in the → instructions for use and/or other relevant → documentation outlining the → intended use and proper use of the AI system and where this use can be foreseen through plausibility reasoning based on knowledge of relevant critical factors (e.g. high likelihood of specific human errors in a specific → use environment or → use context).

#### Explanatory note

Foreseeable misuse always carries the connotation of "reasonably" foreseeable misuse, since it is obviously impossible to anticipate all possible misuses of an AI system once deployed. Foreseeable misuse should be an important consideration when describing and documenting an AI system and should contribute to describing necessary precautions in the → instructions for use.

An evidence-based analysis of foreseeable use may involve surveys with end users (e.g. healthcare professionals) in order to gauge potential scenarios of misuse or may be based on information gathered during → clinical investigations as well as → post-market surveillance activities.

A foreseeable misuse may be, for example, the use of images with other (quality) specifications than those specified in the relevant documentations of the AI system. Such misuse risks that the AI system does not perform as intended and may have severe consequences on the → precision or → accuracy of the output and hence its → safety. Foreseeable misuse does not cover use errors that are due to inadequate → instructions for use or inadequate training of users (i.e. inadequate information).

#### Term relationship:

Related terms:

- Real-world use
- Use error\*

\*) EU MDR's Definition 59 (Article 2; [EU, 2017a](#)) distinguishes use errors from inadequacy of information (see explanatory note above).

## Instructions for use

### Cluster: B.12 Use of AI systems in health and healthcare

#### Concept description

In the broadest sense, instructions for use (IFU) simply mean all relevant explanations, indications, precautions, warnings and information that relate to the use of given product. Instructions for use are a key element for AI systems and in particular AI systems with or supporting a medical purpose.

IFU play an important role in regulatory contexts where a clear delineation of the proper use of an AI system is critical in view of *safety*. IFU can, where required, outline the → use context or requirements in relation to the → use environment.

#### Explanatory note

The EU Regulations on medical devices (MDR) and in vitro diagnostic medical devices (IVDR) define instructions for use as follows: „Instructions for use“ means the information provided by the manufacturer to inform the user of a device's intended purpose and proper use and of any precautions to be taken.“ ([EU, 2017a,b](#))

## Term relationship:

### Related terms:

- Instructions for use
- Real-world use
- Use environment
- Use context

## Real-world use

### Cluster: B.12 Use of AI systems in health and healthcare

#### Concept description

Use of the AI system under conditions and circumstances that can be reasonably expected given the → use context, scenario and the reasonably expectable variability of → use environments. Real-world use goes beyond the optimised conditions designed during validation stages (e.g. → AI system validation (analytical/technical validation, → usability validation, → clinical validation)). While real-world use may differ from use conditions as intended, it does not include alternative use contexts or (foreseeable) misuse.

#### Explanatory note

Conditions and circumstances of real-world use should be given sufficient consideration designing and developing AI systems of sufficient ‘robustness’ (→ risks related to insufficient robustness / resilience). An exploration of potential real-world use scenarios can be undertaken in the context of dedicated → user research during early life cycle stages.

For AI systems used in healthcare, we recommend structured dialogues and surveys of healthcare professionals and, depending on → use context / → use environment, also end users (e.g. patients, caregivers etc.).

#### Term relationship:

### Related terms:

- Intended use

## Use environment

### Cluster: B.12 Use of AI systems in health and healthcare

#### Concept description

Environment in which the AI system is designed to be used and/or in which it is used in practice. The use environment is determined by organisational, operational, technological, infrastructure, legal and human factors.

We define high-level categories of environments as corresponding to the application areas of AI in health (WHO, 2021a): biomedical or clinical research settings, public health institutions, bodies administering health systems and healthcare. In healthcare, there are variety of categories of use environments that correspond to specific settings (e.g. hospitals, clinics, rehabilitation clinics, home care). Importantly,

the actual characteristics of a real-world use environment are shaped by the specific *local* configuration of factors shaping a use environment.

These include:

*Organisation factors:*

- Administrative workflows, reporting lines and clarity of roles
- Quality systems (→ quality culture and risk management)
- Accountability structures (→ accountability structures, attribution of distributed responsibilities), culture, processes and protocols concerning oversight over AI systems (→ ensuring human agency and oversight)

*Technological infrastructure and environment*

- Specifications of IT systems required for the proper operation of the AI system ([Reina & Griesinger, 2024b](#)) within the target use environment
- Interoperability with existing local IT infrastructure (hardware and software) and other products and devices required for the operation of the AI system
- Data availability and processes for data management
- Measures, protocols and process aimed at cybersecurity

See also → risks relating to technical integration and interoperability issues

*Human skills and attitudes*

- Skill set of IT system administrators
- Skill of users (e.g. healthcare professionals, patients, researchers, public health officials)
- User attitudes towards technology in general and specifically AI-enabled technology. This includes automation bias and complacency (→ avoiding automation bias, → avoiding automation complacency) and touches, for healthcare, on the doctor-patient relationship (→ upholding a trustful patient-physician relationship).
- Workload within a given use environment, stress levels and interpersonal dynamics that may impact on the safe use of the AI system.

*Patient-related factors:*

- Patient population characteristics including demographics in a given healthcare use environment
- Incident rate of specific diseases / conditions in the given healthcare environment

*Ethical, legal and regulatory factors*

- Applicable legal and regulatory frameworks, including for data privacy (→ medical privacy / health privacy), → data security, → data protection, as well as → responsiveness, contestability, redress, liability.
- Applicable ethical codes or guidelines (→ AI ethics; → AI principles and ethics guidelines)
- Guidelines, processes and procedures concerning ethical principles relating to → DIGNITY, FREEDOM AND AUTONOMY, in particular those that directly and immediately impact on the use of AI systems in healthcare. These include processes for → ensuring the means for free and informed consent, processes for implementing the → right to know and right not to know and the → right to know if AI system employed.

#### **Explanatory note**

For AI systems used for medical purposes, particular attention should be given to aspects of cybersecurity and end-user skills, competences and understanding (e.g. use within a clinical setting or use for home care, managed by patients or care persons).

## Term relationship:

Related terms:

- Use context
- Upholding a trustful patient-physician relationship

## Use context

### Cluster: B.12 Use of AI systems in health and healthcare

#### Concept description

With use context we mean: the way an AI system is *embedded* or *integrated* within a specific context of use. Determining factors include:

- How the system is intended to *contribute to specified objectives*
- how it is *integrated* and *used within specific processes and workflows* (see also → [use environment](#)) in order to achieve these objectives. In healthcare this may concern clinical workflows and → [clinical pathways](#).
- How its outputs are utilised *in practice* in view of these objectives.

The use context may be proposed in the description of the intended purpose or → [intended use](#) of the AI system. In cases where the use of the AI system is still subject change in the post-deployment space (e.g. in health research), this may not be practicable.

Considerations of use context are important when drafting → [instructions for use](#) and should also be considered in requirement specification for specific → [use requirements](#).

#### Explanatory note

Considerations of use context can benefit from a close dialogue between developer and user, in particular in case of AI systems used for medical purposes, where the use context can be documented in evidence-based → [clinical practice guidelines](#) and/or → [clinical pathways](#).

As an example, consider an AI system for radiological image analysis in healthcare: the → [intended use](#) of an AI system should describe the purpose of the AI system and how to ensure correct outputs (e.g. classification recommendation concerning the likelihood of disease presence/absence with well outlined accuracy values), while the use context should describe how the AI system's output is used within or contributes to broader objectives and how, in view of these objectives, it is embedded in specific processes and procedures, e.g. a specific clinical practice (guideline) for treating a specific disease or condition.

## Term relationship:

Related terms:

- Use environment

## Clinical practice guideline (CPG)

### Cluster: B.12 Use of AI systems in health and healthcare

#### Concept description

'Clinical practice' refers to a "model of practice", typically summarised in evidence-based 'Clinical Practice Guidelines' (CPGs) or simply practice guidelines. These aim at summarising and aggregate medical knowledge based on reliable evidence in order to support physicians in their daily practice in relation to a specific clinical problem, e.g. by gathering evidence for and against the use of a particular diagnostic or therapeutic decision ([Schlieter & Esswein, 2010](#); [Institute of medicine, 2011](#)). CPGs acknowledge that deviations from such codified guidelines may be required due to specific circumstances.

CPGs should be developed in an evidence-based fashion (e.g. based on systematic review(s) of the existing evidence) and should feature "desirable attributes" such as being accurate, accountable, evaluable and able to facilitate confliction resolution and application ([Mayer, 2006](#); [Institute of medicine, 2011](#)).

The Institute of Medicine's committee on standards for developing trustworthy clinical practice guidelines ([Institute of Medicine., 2011](#)) has proposed a catalogue of standards concerning the development of CPGs, including inter alia transparency, management of conflicts of interests (see → **TRUST & TRUSTWORTHINESS – trust issues relation to situations of conflicts of interests**), financial divestment, exclusions of membership concerning a particular CPG (e.g. on grounds of conflicts of interests; see [Guerra-Farfán et al., 2023](#)), composition of expert group, evidence-based tools and how to phrase recommendations. There should also be an external review and a policy and processes for regular updates.

#### Explanatory note

CPGs are not intended to replace clinical freedom (i.e. the → **autonomy** of physicians and patients; → **upholding a trustful patient-physician relationship**). They should be followed in a first tier approach, with deviations being justifiable due to specific circumstances. CPGs can be the basis for step-by-step instructions outlined in → **clinical treatment protocols**. CPGs can also be used for the delineation of a → **clinical pathway**.

#### Term relationship:

Related terms:

- Clinical practice protocol (CPP)

## Clinical practice protocol

### Cluster: B.12 Use of AI systems in health and healthcare

#### Concept description

Clinical practice protocols (CPPs) describe specific steps that should be followed in case of a specific clinical problem. CPPs can be based on workflows that have proven to be effective.

#### Explanatory note

There may be overlaps between CPPs and → **clinical practice guidelines** (CPGs). However, typically CPGs supports the management of a clinical problem by providing a summary of the available evidence, while CPPs typically outline the concrete steps that should be followed once a clinical management decision has been concluded. → **AI-enabled medical device software** (→ **AI-enabled software in a medical device (AI-SIMD)** or → **AI-enabled software as a medical device (AI-SAMD)**) are likely to play an increasingly important role within clinical practice protocols.

- N.B. CPPs should not be confused with “protocols for clinical practice guidelines”, which are protocols for the *development* of clinical practice guidelines, see for instance [Yun et al. \(2023\)](#).

#### Term relationship:

Synonyms:

- Clinical treatment protocol

#### References and further reading

- Hewitt-Taylor J. (2004) Clinical guidelines and care protocols. *Intensive Crit Care Nurs.* 2004 Feb; 20(1):45-52. doi: 10.1016/j.iccn.2003.08.002. PMID: 14726253.
- Scottish NHS (2023) Key Definitions: Decision-Making Support Tools, Clinical Guidelines, Policies, Protocols, Procedures & Care Pathways. Online: <https://rightdecisions.scot.nhs.uk/media/2672/1-key-definitions-decision-making-support-tools.pdf>

## Clinical pathway

#### Cluster: B.12 Use of AI systems in health and healthcare

#### Concept description

Clinical pathways (CPs) are a key instrument for documentation and managing of clinical processes ([Kinsman et al., 2010](#)). CPs document and transfer evidence-based knowledge in practice. CPs are pivotal for clinical quality management processes. CPs may be based on one more → **clinical practices** and associated guidelines.

#### Explanatory note

The use of → **foundation models** and → **generative AI** or large multi-modal models (with the property of generating or synthesizing new content) is currently discussed in regard to their potential usefulness and risks for *inter alia* management of clinical processes or electronic health records (EHR).

#### Term relationship:

Synonyms:

- Care pathway
- Integrate care pathway
- Critical pathway
- Care map

## Applicability and limitations

#### Cluster: B.12 Use of AI systems in health and healthcare

#### Concept description

With applicability or ‘applicability domain’ we refer to the area or areas to which an AI system is *applicable* (e.g. diseases, patient groups, situations, → **use contexts** and → **use environments**) as well as precautions that need to be taken into account under specific circumstances (i.e. *limitations*).

A description of the applicability and limitations of an AI system can draw on its initial design intentions (→ **intended use**) and, more importantly, (real-world) evidence obtained before and after deployment

(e.g. through → model evaluation, → clinical evaluation, → clinical evaluation, → post-market clinical follow-up).

This includes new information such as → generalisability of the AI system, hidden → biases and → usability issues with clear impact on the applicability of the AI system. Updates of applicability and limitation descriptions should be informed by relevant information as it becomes available, e.g. from → post-deployment monitoring activities and → post-market surveillance.

#### Explanatory note

*Applicability* in the context of AI in health medicine, could for instance mean the applicability for supporting clinical decision-making or diagnosis of a specific disease/condition (e.g. a mental health condition) for a specific population or range of patients, while *limitations* would explicitly spell out other diseases/situations or patient parameters to which the AI system may be applicable, albeit with specific limitations (e.g. diagnostic accuracy).

#### Term relationship:

Related term:

- Intended use
- Use context
- Use environment

## **5 Conclusion**

This ontology is a cornerstone of the AI evidence pathway for health aimed to support multi-disciplinary collaboration for the evidence required to build trustworthy AI solutions that can be relied on in healthcare workflows and other health applications.

Trustworthy AI hinges on ethical principles. The ontology is intended to provide an educational resource for unfolding ethical principles into practice-oriented concepts related to the development, assessment, deployment and monitoring of AI in health. It is, to our knowledge, the first attempt of a more systematic and evidence-based exploration of AI ethics and translational concepts (part A) which it connects to fundamental and practice-oriented scientific, clinical, technical, and underpinning ethical, philosophical and societal concepts (part B). The orchestration of these two parts in a modular approach is intended to facilitate community discussion and possible update of this ontology.

We hope that this ontology and the underpinning ‘AI evidence pathway’ will support the emerging concept of AI governance in health and support a forward looking and interconnected pathway of evidence generation that satisfies the needs of various practitioners including AI designers, data scientists, clinicians, regulatory experts and experts focusing on health technology assessment.

## References

### A

1. Abbasgholizadeh Rahimi S et al. (2022) Application of Artificial Intelligence in Shared Decision Making: Scoping Review. JMIR Med Inform. Aug 9;10(8):e36199. Online: doi: 10.2196/36199.
2. Abdallah S et al. (2023) The Impact of Artificial Intelligence on Optimizing Diagnosis and Treatment Plans for Rare Genetic Disorders. Cureus. 2023 Oct 11;15(10):e46860. doi: 10.7759/cureus.46860
3. Abdusalomov AB et al. (2023) Evaluating Synthetic Medical Images Using Artificial Intelligence with the GAN Algorithm. Sensors (Basel). Mar 24;23(7):3440. Online: doi: 10.3390/s23073440
4. Abràmoff MD et al. (2023a) Autonomous artificial intelligence increases real-world specialist clinic productivity in a cluster-randomized trial. npj Digit. Med. 6, 184 (2023). Online: <https://doi.org/10.1038/s41746-023-00931-7>
5. Abràmoff MD et al. (2023b) Considerations for addressing bias in artificial intelligence for health equity. npj Digit. Med. 6, 170 Online: <https://doi.org/10.1038/s41746-023-00913-9>
6. Access partnership (2024). The human cost of AI: Is data labelling creating digital sweatshops? Online: <https://accesspartnership.com/the-human-cost-of-ai-is-data-labelling-creating-digital-sweatshops/>
7. Ada Lovelace Institute (2022) Algorithmic impact assessment: a case study in healthcare. Online: <https://www.adalovelaceinstitute.org/report/algorithmic-impact-assessment-case-study-healthcare/>
8. Ada Lovelace Institute (2024) Safe beyond sale: post-deployment monitoring of AI. Online: <https://www.adalovelaceinstitute.org/blog/post-deployment-monitoring-of-ai/>
9. Adams J (2023) Defending explicability as a principle for the ethics of artificial intelligence in medicine. Med Health Care and Philos 26, 615–623. Online: <https://doi.org/10.1007/s11019-023-10175-7>
10. Adamson AS & Smith A (2018) Machine Learning and Health Care Disparities in Dermatology. JAMA Dermatol. 154(11):1247–1248. Online: doi:10.1001/jamadermatol.2018.2348
11. Addis T (2014) Natural and artificial reasoning. An exploration of modelling human thinking. Springer. Book series: advanced information and knowledge processing, eds: Brahma S et al. Online: <https://link.springer.com/book/10.1007/978-3-319-11286-2>
12. Adebayo J et al. (2018) Sanity checks for saliency maps, Advances in neural information processing systems, vol. 31, arXiv:1810.03292. Online: <https://doi.org/10.48550/arXiv.1810.03292>
13. Afnan MAM et al. (2021) Interpretable, not black-box, artificial intelligence should be used for embryo selection. Hum Reprod Open. 2021 Nov 2;2021(4):hoab040. Online: doi: 10.1093/hropen/hoab040.
14. Agostinelli A et al. (2023) MusicLM: Generating Music From Text. arXiv. Online: <https://doi.org/10.48550/arXiv.2301.11325>
15. Ahmed MI et al. (2023) A Systematic Review of the Barriers to the Implementation of Artificial Intelligence in Healthcare. Cureus. Oct 4;15(10):e46454. Online: doi: 10.7759/cureus.46454
16. AI ethics group (led by VDE / Bertelsmann Stiftung) (2020) From Principles to Practice – An interdisciplinary framework to operationalise AI ethics. Online: <https://www.ai-ethics-im-pact.org/en>
17. Aizenberg E & van den Hoven J (2020) Designing for human rights in AI. Big data & society. July-December: 1-14. Online: DOI: 10.1177/2053951720949566
18. Ajlouni N et al. (2023) Medical image diagnosis based on adaptive Hybrid Quantum CNN. BMC Med Imaging. Sep 14;23(1):126. Online: doi: 10.1186/s12880-023-01084-5

19. Ali S et al., (2023) Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, Volume 99, 101805, ISSN 1566-2535. Online: <https://doi.org/10.1016/j.inffus.2023.101805>.
20. Al Olaimat M & Bozdag S (2024) TA-RNN: an Attention-based Time-aware Recurrent Neural Network Architecture for Electronic Health Records. *arXiv*. Online: <https://doi.org/10.48550/arXiv.2401.14694>
21. Alami H et al. (2020) Artificial Intelligence and Health Technology Assessment: Anticipating a New Level of Complexity. *J Med Internet Res*. Jul 7;22(7):e17707. Online: doi: 10.2196/17707
22. Alan Turing Institute (2024) AI Explainability in practice. AI ethics and governance in practice programme. Online: <https://aiethics.turing.ac.uk/modules/explainability/>
23. Ali S et al., (2023) Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, Volume 99, 101805, ISSN 1566-2535. Online: <https://doi.org/10.1016/j.inffus.2023.101805>.
24. Alfrink K et al. (2023) Contestable AI by Design: Towards a Framework. *Minds & Machines* 33, 613–639 (2023). Online: <https://doi.org/10.1007/s11023-022-09611-z>
25. Algolia (2023) How continuous learning lets machine learning provide increasingly accurate predictions and recommendations. Online: <https://www.algolia.com/blog/ai/how-continuous-learning-lets-machine-learning-provide-increasingly-accurate-predictions-and-recommendations/>
26. Algorithmwatch.org Online: <https://inventory.algorithmwatch.org/>
27. Ali M (2023) Understanding Data Drift and Model Drift: Drift Detection in Python. Online: <https://www.datacamp.com/tutorial/understanding-data-drift-model-drift>
28. Aliouche H (2022) Heuristic decision making in medicine. *News medical life sciences*. Online: <https://www.news-medical.net/health/Heuristic-Decision-Making-in-Medicine.aspx> (Last accessed 2024.08.12).
29. Alowais AA et al. (2023) Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Med Educ*. 2023 Sep 22;23(1):689. Online: doi: 10.1186/s12909-023-04698-z.
30. Amann J et al. (2020) Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak* 20, 310. Online: <https://doi.org/10.1186/s12911-020-01332-6>
31. Amazon Web Services (AWS) (2024) Best practices for retiring applications before decommissioning infrastructure. AWS prescriptive guidance. Online: <https://docs.aws.amazon.com/pdfs/prescriptive-guidance/latest/migration-app-retirement-best-practices/migration-app-retirement-best-practices.pdf>
32. American Medical Association, AMA (2024) Augmented Intelligence in Health Care H-480.940. Online: <https://policysearch.ama-assn.org/policyfinder/detail/augmented%20intelligence?uri=%2FAMADoc%2FHOD.xml-H-480.940.xml>
33. American Medical Association (2024) Augmented intelligence in medicine. Online: <https://www.ama-assn.org/practice-management/digital/augmented-intelligence-medicine> (last accessed 2024.08.05)
34. Anabheri R (2024) Data integrity in AI: combating deceptive AI-generated outputs. Online: <https://www.forbes.com/councils/forbestechcouncil/2024/06/20/data-integrity-in-ai-combating-deceptive-ai-generated-outputs/> (Last accessed: 2024.08.13).
35. Ananny M & Crawford K (2018) Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), 973-989. <https://doi.org/10.1177/1461444816676645>
36. Andrews C (2025) European Commission withdraws AI Liability Directive from consideration. <https://iapp.org/news/a/european-commission-withdraws-ai-liability-directive-from-consideration>

37. Aquino YSJ et al. (2022) Professional Perspectives on the Impact of Healthcare Artificial Intelligence on Clinical Roles and Skills. Online: <http://dx.doi.org/10.2139/ssrn.4129747> (Last accessed: 2024.09.18)
38. Arbib MA (1987) Computer and brain. In: Encyclopedia of neuroscience. Vol I. pp. 269-270. Birkhäuser, Boston.
39. Arnold MH (2021) Teasing out Artificial Intelligence in Medicine: An Ethical Critique of Artificial Intelligence and Machine Learning in Medicine. *J Bioeth Inq*. Mar;18(1):121-139. Online: doi: 10.1007/s11673-020-10080-1. Epub 2021 Jan 7.
40. Arora A et al. (2023) Risk and the future of AI: Algorithmic bias, data colonialism, and marginalization, Information and Organisation. Volume 33, Issue 3, 100478. Online: <https://doi.org/10.1016/j.infoandorg.2023.100478> or <https://www.sciencedirect.com/science/article/pii/S1471772723000325>
41. Arora A & Arora (2022) Generative adversarial networks and synthetic patient data: current challenges and future perspectives. *Future Healthc J*. Jul;9(2):190-193. Online: doi: 10.7861/fhj.2022-0013.
42. Ashcroft R (2002) What is clinical effectiveness? *Stud. Hist. Phil. Biol. & Biomed. Sci* 33:219-233. Online: <https://www.sciencedirect.com/science/article/pii/S0039368102000201>
43. Atasever S et al. (2023) A comprehensive survey of deep learning research on medical image analysis with focus on transfer learning. *Clin Imaging*. Feb;94:18-41. Online: doi: 10.1016/j.clinimag.2022.11.003.
44. Awan A (2024) What is algorithmic bias? Online: <https://www.datacamp.com/blog/what-is-algorithmic-bias>
45. Ayana G et al. (2024) Multistage transfer learning for medical images. *Artif Intell Rev* 57, 232. Online: <https://doi.org/10.1007/s10462-024-10855-7>
46. Ayilara OF et al. (2023) Generating synthetic data from administrative health records for drug safety and effectiveness studies. *Int J Popul Data Sci*. 2023 Nov 27;8(1):2176. Online: doi: 10.23889/ijpds.v8i1.2176.

## B

47. Babushkina D (2023) Are we justified attributing a mistake in diagnosis to an AI diagnostic system? *AI Ethics* 3, 567-584. <https://doi.org/10.1007/s43681-022-00189-x>
48. Badimon L, Padro T & Vilahur G (2023) Chapter 27 - Moving from reactive to preventive medicine. Editor(s): Paulo J. Oliveira, João O. Malva, Aging, Academic Press, pages 663-681. Online: <https://doi.org/10.1016/B978-0-12-823761-8.00003-3>.
49. Bagnis A et al., (2021) Facing up to bias in healthcare: The influence of familiarity appearance on hiring decisions. *Applied Cognitive Psychology*. 35(6): 1585-1591. Online: <https://onlinelibrary.wiley.com/doi/full/10.1002/acp.3873>
50. Baier L et al. (2019) Challenges in the deployment and operation of machine learning in practice. Twenty-Seventh European Conference on Information Systems (ECIS2019), Stockholm-Uppsala, Sweden Online: [https://www.researchgate.net/publication/332996647\\_CHALLENGES\\_IN\\_THE\\_DEPLOYMENT\\_AND\\_OPERATION\\_OF\\_MACHINE\\_LEARNING\\_IN\\_PRACTICE](https://www.researchgate.net/publication/332996647_CHALLENGES_IN_THE_DEPLOYMENT_AND_OPERATION_OF_MACHINE_LEARNING_IN_PRACTICE)
51. Bajwa J et al. (2021) Artificial intelligence in healthcare: transforming the practice of medicine. *Future Healthc J*. Jul;8(2):e188-e194. Online: doi: 10.7861/fhj.2021-0095.
52. Baker JD (2024) 5 Ways to Cut AI Energy Consumption. Online: <https://builtin.com/articles/ai-energy-consumption>
53. Baowaly MK et al. (2019) Synthesizing electronic health records using improved generative adversarial networks. *J Am Med Inform Assoc*. Mar 1;26(3):228-241. Online: doi: 10.1093/jamia/ocz142

54. Baracaldo N & Xu R (2022) Protecting Against Data Leakage in Federated Learning: What Approach Should You Choose? In: "Federated learning: a comprehensive overview of methods and applications" eds: Ludwig H & Baracaldo N. Springer. Online: [https://link.springer.com/chapter/10.1007/978-3-030-96896-0\\_13](https://link.springer.com/chapter/10.1007/978-3-030-96896-0_13)
55. BCLP LLP (2024) US state-by-state AI legislation snapshot. Online: <https://www.bclplaw.com/en-US/events-insights-news/us-state-by-state-artificial-intelligence-legislation-snapshot.html>
56. Beattie P & Nelson R (2006) Clinical prediction rules: what are they and what do they tell us? Aust J Physiother. 52(3):157-63. Online: doi: 10.1016/s0004-9514(06)70024-1
57. Beauchamp TL & Childress JF (1979; republished 2001). Principles of biomedical ethics. Oxford University Press, USA. Available at Google books: [https://books.google.it/books/about/Principles\\_of\\_Biomedical\\_Ethics.html?id=\\_14H7M0w1o4C&redir\\_esc=y](https://books.google.it/books/about/Principles_of_Biomedical_Ethics.html?id=_14H7M0w1o4C&redir_esc=y)
58. Beauchamp TL & Rauprich O (2016). Principlism. In: ten Have, H. (eds) Encyclopedia of Global Bioethics. Springer, Cham. Online: [https://doi.org/10.1007/978-3-319-09483-0\\_348](https://doi.org/10.1007/978-3-319-09483-0_348)
59. Bedi S et al. (2025) Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review. JAMA. 2025 Jan 28;333(4):319-328. doi: 10.1001/jama.2024.21700
60. Begg et al. (1996). Improving the Quality of Reporting of Randomized Controlled Trials. The CONSORT Statement. Online: <https://doi.org/10.1001/jama.276.8.637>
61. Beisbart C & Räz T (2022) Philosophy of science at sea: Clarifying the interpretability of machine learning. Philosophy Compass, 17(6), e12830. <https://doi.org/10.1111/phc3.12830>
62. Berger FR (1979; online 2020) Canadian Journal of Philosophy Supplementary Volume, Vol. 5: New Essays on John Stuart Mill and Utilitarianism, 1979, pp. 115 – 136. Online: DOI: <https://doi.org/10.1080/00455091.1979.10717097>
63. Bernardi FA et al. (2023). Data Quality in Health Research: Integrative Literature Review. J Med Internet Res. Oct 31;25:e41446. Online: doi: 10.2196/41446
64. Bialek J (2024) Understanding Data Drift: Impact on Machine Learning Model Performance. Online: <https://www.nannyml.com/blog/types-of-data-shift> (Last accessed: 2024.09.02)
65. Bibbins-Domingo K et al. (2024) The 2024 Revision to the Declaration of Helsinki: Modern Ethics for Medical Research. JAMA. Available online: doi:10.1001/jama.2024.22530
66. Bikku, T. (2020) Multi-layered deep learning perceptron approach for health risk prediction. J Big Data 7, 50. Online: <https://doi.org/10.1186/s40537-020-00316-7>
67. Biran O & Cotton CV (2017) Explanation and Justification in Machine Learning : A Survey. In IJCAI-17 workshop on explainable AI (XAI).
68. Bishop C (2016) Pattern Recognition and Machine Learning. Eds: Jordan, Kleinberg, Schölkopf. Springer, New York. Online available: <https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf>
69. Biswas A (2024) A comprehensive review of explainable AI for disease diagnosis, Array, Volume 22, 100345. Online: <https://doi.org/10.1016/j.array.2024.100345>
70. Bitkina OV et al. (2020) Usability and user experience of medical devices: An overview of the current state, analysis methodologies, and future challenges. International Journal of Industrial Ergonomics, Volume 76, 102932, ISSN 0169-8141, <https://doi.org/10.1016/j.ergon.2020.102932>
71. Black CL (2023) Heuristic evaluation of portable pulse oximeters for domiciliary use: Implications for its use in assessing medical device usability. Smart Health, Volume 27. Online: <https://doi.org/10.1016/j.smhl.2022.100357>
72. Black A & van Nederpelt P (2020) Dimensions of data quality – research paper. Online: <https://www.dama-nl.org/wp-content/uploads/2020/09/DDQ-Dimensions-of-Data-Quality-Research-Paper-version-1.2-d.d.-3-Sept-2020.pdf>

73. Block HD (1962) The perceptron: a model for brain functioning. *Rev. Mod. Phys.* 34, 123. Online: DOI: <https://doi.org/10.1103/RevModPhys.34.123>
74. Block HD (1970) A review of “Perceptrons: An introduction to computational geometry”. *Information and Control*, Volume 17, Issue 5, Pages 501-522. Online: [https://doi.org/10.1016/S0019-9958\(70\)90409-2](https://doi.org/10.1016/S0019-9958(70)90409-2).
75. Boehm KM et al. (2022) Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. *Nat Cancer*. 2022 Jun;3(6):723-733. Online: doi: 10.1038/s43018-022-00388-9
76. Bommasani R et al. (2021) On the opportunities and risks of foundation models, 2021. Online: <http://arxiv.org/abs/2108.07258>.
77. Borah A, Nath (2018). Identifying risk factors for adverse diseases using dynamic rare association rule mining. *Expert Syst Appl.* 113:233–63. Online: <https://doi.org/10.1016/j.eswa.2018.07.010>
78. Borsci S & David LZ (2020) Chapter 117 – Human factors and system thinking for medical device. Editor(s): Ernesto Iadanza. *Clinical Engineering Handbook* (Second Edition). Academic Press. Pages 829-831. ISBN 9780128134672. Online: <https://doi.org/10.1016/B978-0-12-813467-2.00118-8>.
79. Borsci S (2018) Designing medical technology for resilience: integrating health economics and human factors approaches. *Expert Rev Med Devices*. Jan; 15(1):15-26. Online: doi:10.1080/17434440.2018.1418661
80. Borys K et al. (2023) Explainable AI in medical imaging: An overview for clinical practitioners - Saliency-based XAI approaches. *Eur J Radiol*. May;162:110787. Online: doi: 10.1016/j.ejrad.2023.110787
81. Bossuyt et al. (2015). STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. On-line: <doi.org/10.1136/bmj.h5527>
82. Bozinovski S (2020) Reminder of the First Paper on Transfer Learning in Neural Networks, 1976. *Informatica* 44, 291–302. Online: <https://doi.org/10.31449/inf.v44i3.2828>
83. Brady AP & Neri E (2020) Artificial Intelligence in Radiology-Ethical Considerations. *Diagnostics* (Basel). Apr 17;10(4):231. Online: doi: 10.3390/diagnostics10040231.
84. Brahnam S (2006) Gendered bots and bot abuse. In Antonella de Angeli, Sheryl Brahnam, Peter Wallis, & Peter Dix (Eds.), *Misuse and abuse of interactive technologies* (pp. 1–4). Montreal: ACM. Online: [https://www.researchgate.net/publication/220054517\\_Special\\_issue\\_on\\_the\\_abuse\\_and\\_misuse\\_of\\_social\\_agents](https://www.researchgate.net/publication/220054517_Special_issue_on_the_abuse_and_misuse_of_social_agents)
85. Braitenberg V & Schüz A (1991) Anatomy of the cortex – statistics and geometry. Springer, Berlin. Online (paywall): <https://link.springer.com/book/10.1007/978-3-662-02728-8>
86. Breiman L (2001) Statistical Modeling: The Two Cultures. *Statistical Science* 16(3), 199-231. Online: <https://projecteuclid.org/journals/statistical-science/volume-16/issue-3/Statistical-Modeling--The-Two-Cultures-with-comments-and-a/10.1214/ss/1009213726.full>
87. Braun MN et al. (2021) A meta-analysis of Libet-style experiments. *Neuroscience & Biobehavioral Reviews*. Volume 128: 182-198. Online: <https://doi.org/10.1016/j.neubiorev.2021.06.018>
88. Brothers KB & Rothstein MA (2015) Ethical, legal and social implications of incorporating personalized medicine into healthcare. *Per Med*.12(1):43-51. Online: doi: 10.2217/pme.14.65.
89. Brown PR (2009) The phenomenology of trust: A Schutzian analysis of the social construction of knowledge by gynaec-oncology patients. *Health, Risk & Society*, 11(5), 391–407. Online: <https://doi.org/10.1080/13698570903180455>
90. Brownlee J (2020) Difference Between Algorithm and Model in Machine Learning. Online: <https://machinelearningmastery.com/difference-between-algorithm-and-model-in-machine-learning/>

91. Brunel N & van Rossum MCW (2007) Quantitative investigations of electrical nerve excitation treated as polarization. Translation of the paper by Lapicque published in 1907 in French. Biol Cybern 97, 341–349. Online: <https://doi.org/10.1007/s00422-007-0189-6>
92. Burau V (2021) Personalised medicine and the state: A political discourse analysis. Health Policy. Jan;125(1):122-129. Online: doi: 10.1016/j.healthpol.2020.10.005
93. Burches E & Burches M (2020) Efficacy, Effectiveness and Efficiency in the Health Care: The Need for an Agreement to Clarify its Meaning. Int Arch Public Health Community Med 4:035. doi.org/10.23937/2643-4512/1710035
94. Burkart N & Huber MF (2021) A survey on the explainability of supervised machine learning. Journal of Artificial Intelligence Research, vol. 70, pp. 245– 317, 2021. Online: <https://dl.acm.org/doi/10.1613/jair.1.12228>
95. Busch F et al. (2024) Navigating the European Union Artificial Intelligence Act for Healthcare. npj Digit. Med. 7, 210 (2024). Online: <https://doi.org/10.1038/s41746-024-01213-6>
96. Bzdok D, Altman N, Krzywinski M (2018) Statistics versus machine learning. Nature Methods, Vol.15(4):233-234. Online: <https://doi.org/10.1038/nmeth.4642>

## C

97. Campolo A et al. (2017) AI NOW 2017 Report. Eds: Selbst A & Baracas S. Online: [https://as-sets.ctfas-sets.net/8wprhhvnpfc0/1A9c3ZTC7a2KEYM64Wsc2a/8636557c5fb14f2b74b2be64c3ce0c78/\\_AI\\_Now\\_Institute\\_2017\\_Report.pdf](https://as-sets.ctfas-sets.net/8wprhhvnpfc0/1A9c3ZTC7a2KEYM64Wsc2a/8636557c5fb14f2b74b2be64c3ce0c78/_AI_Now_Institute_2017_Report.pdf)
98. Cai L & Zhu Y (2015) The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal 14(0), p. 2. Online: <https://doi.org/10.5334/dsj-2015-002>
99. Calo R (2018) Artificial intelligence policy: a primer and roadmap. University of Bologna Law Review Vol. 3(2):180-218. Online: <https://doi.org/10.6092/issn.2531-6133/8670>
100. Capes, J.A. (2017) Freedom with Causation. Erkenn 82, 327–338. Online: <https://doi.org/10.1007/s10670-016-9819-5>
101. Carnap, Rudolf (1950). Logical foundations of probability. Chicago]: Chicago University of Chicago Press. See also: Stanford encyclopedia of philosophy. Entry : supplement to Rudolf Carnap. Online: <https://plato.stanford.edu/entries/carnap/methodology.html>
102. Carreira-Perpinan, MA et al. (2005) A computational model for the development of multiple maps in primary visual cortex. Cereb. Cortex 15, 1222–1233 (2005). Online: <https://doi.org/10.1093/cercor/bhi004>
103. Čartolovni A, Tomičić A, Lazić Mosler E (2022) Ethical, legal, and social considerations of AI-based medical decision-support tools: A scoping review. Int J Med Inform. May;161:104738. doi: 10.1016/j.ijmedinf.2022.104738. Epub 2022 Mar 14.
104. Cath C (2018) Governing artificial intelligence: ethical, legal and technical opportunities and challenges. Philos Trans A Math Phys Eng Sci. Oct 15;376(2133):20180080. Online: doi: 10.1098/rsta.2018.0080
105. Caton S & Haas C (2020) Fairness in Machine Learning: A Survey. arXiv:2010.04053. Online: <https://doi.org/10.48550/arXiv.2010.04053>
106. Ceresa M et al. (2024) Retrieval Augmented Generation Evaluation for Health Documents. arXiv. Online: <https://doi.org/10.48550/arXiv.2505.04680>
107. Ceresa M, Comte V, Reina V & Griesinger CB (2025) Generative AI in healthcare. In European Commission – Joint Research Centre (2025) Generative AI outlook report: exploring the intersection of technology, policy and society. Online: doi: 10.2760/1109679
108. CERNA (2018) L'alliance des sciences et technologies du numérique (Allistene) Research ethics in machine learning. Online: <https://hal.science/ALLISTENE-CERNA/hal-01724307v1>

109. Challen R et al. (2019) Artificial intelligence, bias and clinical safety. *BMJ Qual Saf.* Mar;28(3):231–237. doi: 10.1136/bmjqqs-2018-008370. Epub 2019 Jan 12.
110. Chan AW et al. (2013) Text SPIRIT 2013 statement: defining standard protocol items for clinical trials. Online: <https://doi.org/10.7326/0003-4819-158-3-201302050-00583>
111. Chandra K et al. (2019) Gradient descent: the ultimate optimizer. *arXiv.* Online: <https://arxiv.org/pdf/1909.13371>
112. Char DS et al. (2018) Implementing Machine Learning in Health Care - Addressing Ethical Challenges. *N Engl J Med.* Mar 15;378(11):981–983. Online: doi: 10.1056/NEJMp1714229
113. Char DS et al. (2020) Identifying Ethical Considerations for Machine Learning Healthcare Applications. *Am J Bioeth.* Nov;20(11):7–17. Online: doi: 10.1080/15265161.2020.1819469
114. Charpignon ML et al. (2023) Critical Bias in Critical Care Devices. *Crit Care Clin.* Oct;39(4):795–813. Online: doi: 10.1016/j.ccc.2023.02.005. Epub 2023 Mar 27.
115. Chassaigne H et al., (2024) Interrogating the research and development pipeline of artificial intelligence (AI) in health: diagnosis and prediction-based diagnosis. Online: doi:10.2760/8612579
116. Chassaigne H et al (in preparation for publication in 2025) Interrogating the research and development pipeline of AI in health: key attributes and reporting by researchers
117. Chato L & Regentova E (2023) Survey of Transfer Learning Approaches in the Machine Learning of Digital Health Sensing Data. *J Pers Med.* Dec 12;13(12):1703. Online: doi: 10.3390/jpm13121703
118. Chen et al. (2023) How Is ChatGPT's Behavior Changing over Time? Online: arXiv:2302.00487.
119. Chen X (2024) Algorithmic proxy discrimination and its regulations, *Computer Law & Security Review*, Volume 54, 106021. Online: <https://doi.org/10.1016/j.clsr.2024.106021>
120. Chin MH (2023) Guiding Principles to Address the Impact of Algorithm Bias on Racial and Ethnic Disparities in Health and Health Care. *JAMA Netw Open.* Dec 1;6(12):e2345050. doi: 10.1001/jamanetworkopen.2023.45050.
121. Chiodo S (2022) Human autonomy, technological automation (and reverse). *AI & SOCIETY* (2022) 37:39–48. Online: <https://doi.org/10.1007/s00146-021-01149-5>
122. CIOMS - Council for International Organisations of Medical Sciences (2002) International Ethical Guidelines for Biomedical Research Involving Human Subjects. Online: [https://cioms.ch/wp-content/uploads/2021/03/International\\_Ethical\\_Guidelines\\_for\\_Biomedical\\_Research\\_Involving\\_Human\\_Subjects\\_2002.pdf](https://cioms.ch/wp-content/uploads/2021/03/International_Ethical_Guidelines_for_Biomedical_Research_Involving_Human_Subjects_2002.pdf)
123. Chirag S (2023) Types of Sampling in Machine Learning. In 'machine learning newsletter'. Online: <https://www.linkedin.com/pulse/types-sampling-machine-learning-chirag-subramanian-hnsoc/>
124. ChosunBiz news (2024) Korea passes AI Basic Act, second globally, enhancing national AI competitiveness. Online: <https://biz.chosun.com/en/en-it/2024/12/26/66W2Z3RX6FE7FMPXMR73T26SKY/>
125. Chuen-Kai S. et al. (2015) Transfer representation learning for medical image analysis. *Annu Int Conf IEEE Eng Med Biol Soc.* Aug;2015:711–4. Online: doi: 10.1109/EMBC.2015.7318461.
126. Cisco Data Center Networking Solutions: Addressing the Challenges of AI/ML Infrastructure. Online: <https://www.cisco.com/c/en/us/td/docs/dcn/whitepapers/cisco-addressing-ai-ml-network-challenges.html> (Last accessed: 2024.11.14)
127. Clarke R (2019) Principles and business processes for responsible AI. *Comput Law Secur Rev* 35:410–422. Online: <https://doi.org/10.1016/j.clsr.2019.04.007>
128. Cline HT (2023) Activity-dependent Organisation of Topographic Neural Circuits, *Neuroscience*, Volume 508: 3–18. Online: <https://doi.org/10.1016/j.neuroscience.2022.11.032>
129. Clouser KD & Gert B. A (1990) A critique of principlism. *J Med Philos.* Apr;15(2):219–36. Online: doi: 10.1093/jmp/15.2.219.

130. Clusmann J et al. (2023) The future landscape of large language models in medicine. *Commun Med* 3, 141 (2023). <https://doi.org/10.1038/s43856-023-00370-1>
131. Coglianese C (2020) The Law and Economics of Risk Regulation. University of Pennsylvania, Institute for Law & Economics Research Paper No. 20-18, 9. Online: [https://scholarship.law.upenn.edu/faculty\\_scholarship/2157/](https://scholarship.law.upenn.edu/faculty_scholarship/2157/)
132. Coglianese C (2024) A People-and-Processes Approach to AI Governance. Online: The Regulatory Review (8. January 2024). Online: <https://www.theregreview.org/2024/01/08/coglianese-a-people-and-processes-approach-to-ai-governance/>
133. Collins GS et al. (2015) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. Online: <https://doi.org/10.1186/s12916-014-0241-z>
134. Collins GS et al. (2021) Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*. 2021 Jul 9;11(7):e048008. Online: doi: 10.1136/bmjopen-2020-048008 or <https://bmjopen.bmj.com/content/11/7/e048008>
135. Collins GS et al. (2024) TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. Online: <https://doi.org/10.1136/bmj-2023-078378>
136. ComplianceAspekte (2024) PIA or DPIA: What's the Difference? Online: <https://compliance-aspekte.de/en/blog/blog-pia-or-dpia/>
137. Confluent (2024) How Developers Can Use Generative AI to Improve Data Quality. Online: <https://www.confluent.io/blog/how-developers-can-use-generative-ai-to-improve-data-quality/> (last accessed 2024.12.16)

### *Council of Europe*

138. Council of Europe (1950, entering into force 1953). European convention on human rights. Online: <https://www.coe.int/en/web/human-rights-convention>
139. Council of Europe (1981) Convention for the protection of individuals with regard to automatic processing of personal data. ETS No. 108. Modernised convention of 2018: Online: <https://www.coe.int/en/web/data-protection/convention108-and-protocol>
140. Council of Europe (1997a) Oviedo convention on human rights and biomedicine (ETS No 164). Online: <https://www.coe.int/en/web/bioethics/oviedo-convention>
141. Council of Europe (1997b) Explanatory Report to the 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 ("Oviedo Convention"). Online: <https://www.coe.int/en/web/bioethics/oviedo-convention>
142. Council of Europe (2005) Additional Protocol to the Convention on Human Rights and Biomedicine, concerning Biomedical Research (CETS No. 195). Online: <https://www.coe.int/en/web/conventions/full-list?module=treaty-detail&treatynum=195>
143. Council of Europe (2014) Human rights themes, entry "Discrimination and Intolerance". Online <https://www.coe.int/en/web/compass/discrimination-and-intolerance>
144. Council of Europe (2018) Addressing the impact of algorithms on human rights. Strasbourg: Council of Europe' 2019. Online: <https://rm.coe.int/draft-recommendation-of-the-committee-of-ministers-to-states-on-the-hu/168095eecf>
145. Council of Europe – Parliamentary Assembly (2020a) Report 15154 Artificial intelligence in health care: medical, legal and ethical challenges ahead. Online: [https://www.eerstekamer.nl/bijlage/20201105/artificial\\_intelligence\\_in\\_health/document3/f=/vldiey9antv3.pdf](https://www.eerstekamer.nl/bijlage/20201105/artificial_intelligence_in_health/document3/f=/vldiey9antv3.pdf)
146. Council of Europe (2020b) CDDH comments on the Parliamentary Assembly Recommendation: Artificial intelligence in health care: medical, legal and ethical challenges ahead. Online:

- <https://rm.coe.int/recommandation-2185-2020-artificial-intelligence-in-health-care-medica/1680a2dcfa>
147. Council of Europe, Council of Ministers (2022) Reply to Recommendation Artificial intelligence in health care: medical, legal and ethical challenges ahead. Doc. 15508 | 25 April 2022. Online: <https://pace.coe.int/en/files/29935/html>
  148. Council of Europe (2024a) Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law. Online: <https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence>
  149. Council of Europe (2024b) Explanatory report to the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law. Online: <https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence>
  150. Council of Europe (2024c) Council of Europe's website on "AI initiatives". Online <https://www.coe.int/en/web/artificial-intelligence/national-initiatives>. The site allows accessing a non-exhaustive data collection on AI documents. Online: [https://docs.google.com/spreadsheets/d/1mu2brATV\\_fgd5MRGft2ASOFepAI1pivwhGm0VCT22\\_U/edit?qid=0#gid=0](https://docs.google.com/spreadsheets/d/1mu2brATV_fgd5MRGft2ASOFepAI1pivwhGm0VCT22_U/edit?qid=0#gid=0)
  151. Covert Q et al. (2020) Towards a Triad for Data Privacy, Proceedings of the 53rd Hawaii International Conference on System Sciences. Online: <https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/5486a250-cc3c-4227-a752-7d08378afbd/content>
  152. Cowan D (2016) Sampling. In: The science of machine learning and AI. Online: <https://www.ml-science.com/sampling>
  153. Cowley LE et al. (2019) Methodological standards for the development and evaluation of clinical prediction rules: a review of the literature. Diagn Progn Res. Aug 22;3:16. Online: doi: 10.1186/s41512-019-0060-y.
  154. Craver CF (2025) Mechanistic explanation. In: Open encyclopedia of cognitive science. Online: <https://oecs.mit.edu/pub/vqigt1aq/release/1>
  155. Crigger E et al. (2022) Trustworthy Augmented Intelligence in Health Care. J Med Syst. 2022 Jan 12;46(2):12. doi: 10.1007/s10916-021-01790-z.
  156. Cutillo CM et al. (2020) Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. npj Digit. Med. 3, 47. Online: <https://doi.org/10.1038/s41746-020-0254-2>

## D

157. Dallanoce F (2022) Explainable AI: A Comprehensive Review of the Main Methods. Online: Medium January 4. Online: <https://medium.com/mlearning-ai/explainable-ai-a-complete-summary-of-the-main-methods-a28f9ab132f7> (Last accessed: 2024.08.16)
158. Danesh J et al. (1999) Postcodes as useful markers of social class: population based study in 26 000 British households. BMJ. Mar 27;318(7187):843-4. doi: 10.1136/bmj.318.7187.843.
159. Darling K (2016) Extending legal protection to social robots: The effect of anthropomorphism, empathy, and violent behavior towards robotic objects. In R. Calo, A. M. Froomkin, & I. Kerr (Eds.), Robot law (pp. 213–234). Cheltenham: Edward Elgar. Online: <https://doi.org/10.4337/9781783476732.00017>
160. Data Ethics EU (2024) - Schau S (2023) The Digital Extraction of Surplus Value from the Global South. Online: <https://dataethics.eu/the-digital-extraction-of-surplus-value-from-the-global-south/>
161. DataMeaning (2024) Comprehensive Guide to Data Governance for Ensuring Data Privacy: Why It Truly Matters. Online: <https://datameaning.com/2024/05/15/data-privacy-governance/>
162. Datta A et al. (2017) Proxy non-discrimination in data-driven systems. arXiv:1707.08120. Online: <https://doi.org/10.48550/arXiv.1707.08120>

163. Davies B, Savulescu J (2021) The Right Not to Know: some steps towards a compromise. *Ethical Theory Moral Pract.* 2021 Mar;24(1):137-150. Online: doi: 10.1007/s10677-020-10133-9. Epub 2020 Oct 29.
164. De Oliveira R et al. (2024) Improving Energy Efficiency in Federated Learning Through the Optimisation of Communication Resources Scheduling of Wireless IoT Networks. arXiv:2408.01286. Online: <https://doi.org/10.48550/arXiv.2408.01286>
165. De Silva D & Alahakoon D (2021) An artificial intelligence life cycle: from conception to production. arXiv:2108.13861v1. Online: <https://doi.org/10.48550/arXiv.2108.13861>
166. DeCamp M & Lindvall C (2020) Latent bias and the implementation of artificial intelligence in medicine. *J Am Med Inform Assoc.* Dec 9;27(12):2020-2023. Online: doi: 10.1093/jamia/ocaa094.
167. Dechter R (1986) Learning while searching in constraint-satisfaction-problems. *AAAI'86: Proceedings of the Fifth AAAI National Conference on Artificial Intelligence.* Pages 178 – 183. Online: <https://cdn.aaai.org/AAAI/1986/AAAI86-029.pdf>
168. The DECIDE-AI Steering Group (2021) DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nat Med* 27, 186–187 (2021). Online: <https://doi.org/10.1038/s41591-021-01229-5>
169. Deck L et al. (2024) A critical survey on fairness benefits of explainable AI. arXiv:2310.13007v6. Online: <https://arxiv.org/pdf/2310.13007v6.pdf> (Last accessed: 2024.08.20)
170. DeepAI machine learning glossary. Online <https://deepai.org/definitions>
171. Deley T & Dubois E (2020) Assessing Trust Versus Reliance for Technology Platforms by Systematic Literature Review. *Social Media + Society*, 6(2). Online: <https://doi.org/10.1177/2056305120913883>
172. Deloitte (2022) Trustworthy AI – bridging the ethics gap surrounding AI. Online information: <https://www2.deloitte.com/us/en/pages/deloitte-analytics/solutions/ethics-of-ai-framework.html>. Framework is behind paywall.
173. Delpierre C & Lefèvre 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.* Feb 21;8:1112159. Online: doi: 10.3389/fsoc.2023.1112159
174. Dhade P & Shirke P (2023) Federated Learning for Healthcare: A Comprehensive Review. *Engineering Proceedings* 59, no. 1: 230. Online: <https://doi.org/10.3390/engproc2023059230>
175. Di Mattia P (2008) Ethical Principles. In: Kirch, W. (eds) *Encyclopedia of Public Health*. Springer, Dordrecht. [https://doi.org/10.1007/978-1-4419-1033-4\\_1033](https://doi.org/10.1007/978-1-4419-1033-4_1033)
176. Dias AM (2016) Commentary: Free Will and Neuroscience: From Explaining Freedom Away to New Ways of Operationalizing and Measuring It. *Front. Hum. Neurosci.*, 19 October. Sec. Cognitive Neuroscience. Volume 10. Online: <https://doi.org/10.3389/fnhum.2016.00509>
177. Digital Catapults (2023) Ethics framework. Online: [https://www.digitcatapult.org.uk/wp-content/uploads/2023/06/DC\\_AI\\_Ethics\\_Framework-2021.pdf](https://www.digitcatapult.org.uk/wp-content/uploads/2023/06/DC_AI_Ethics_Framework-2021.pdf)
178. Dorsey ER & Ritzer G (2016) The McDonaldization of Medicine. *JAMA Neurol.* Jan;73(1):15-6. doi: 10.1001/jamaneurol.2015.3449.
179. Dunn J. et al. (2021) Comparing interpretability and explainability for feature selection. arXiv: 2105.05328v1. Online: <https://doi.org/10.48550/arXiv.2105.05328>
180. Duran LDD (2021) Deskilling of medical professionals: an unintended consequence of AI implementation? *Giornale di Filosofia.* 2021. 2. Online: <https://mimesisjournals.com/ojs/index.php/giornale-filosofia/article/view/1691/1342>
181. Dutch Ministry of the Interior and Kingdom Relations (2022) Impact Assessment - Fundamental rights and algorithms. Online: <https://www.government.nl/documents/reports/2022/03/31/impact-assessment-fundamental-rights-and-algorithms>

**E**

182. Ebers M et al. (2021) The European Commission's Proposal for an Artificial Intelligence Act—A Critical Assessment by Members of the Robotics and AI Law Society (RAILS). *J.* 2021 4(4):589–603. Online: <https://doi.org/10.3390/j4040043>
183. Ebers M (2024a) Therapy without Therapists: Human–Robot Interaction under the EU Medical Device Regulation and the Artificial Intelligence Act. In: Barfield W, Weng Y-H, Pagallo U, eds. *The Cambridge Handbook of the Law, Policy, and Regulation for Human–Robot Interaction*. Cambridge Law Handbooks. Cambridge University Press; 2024:724–752. Online (paywall): <https://www.cambridge.org/core/books/abs/cambridge-handbook-of-the-law-policy-and-regulation-for-humanrobot-interaction/therapy-without-therapists/BE9960C4478AFBC27CE17A6ECC71B0D2>
184. Ebers M (2024b) Truly Risk-Based Regulation of Artificial Intelligence - How to Implement the EU's AI Act. Online: <http://dx.doi.org/10.2139/ssrn.4870387>
185. Edwards L & Veale M (2017) Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For. *16 Duke Law & Technology Review* 18 (2017), Available at SSRN: <https://ssrn.com/abstract=2972855> or <http://dx.doi.org/10.2139/ssrn.2972855>
186. Ehrlinger L & Wöß W (2022) A Survey of Data Quality Measurement and Monitoring Tools. *Front. Big Data, Sec. Data mining and management*. Volume 5 – 2022. Online: <https://doi.org/10.3389/fdata.2022.850611>
187. Ehrmann DE et al. (2023) Making machine learning matter to clinicians: model actionability in medical decision-making. *npj Digit. Med.* 6, 7. Online: <https://doi.org/10.1038/s41746-023-00753-7>
188. Eide AW et al. (2016) Human-Machine Networks: Towards a Typology and Profiling Framework. In: Kurosu, M. (eds) *Human-Computer Interaction. Theory, Design, Development and Practice*. HCI 2016. Lecture Notes in Computer Science(), vol 9731. Springer, Cham. [https://doi.org/10.1007/978-3-319-39510-4\\_2](https://doi.org/10.1007/978-3-319-39510-4_2)
189. Elahi B (2022) Safety risk management for medical devices. Elsevier. Online: <https://www.sciencedirect.com/book/9780323857550/safety-risk-management-for-medical-devices#book-info> (See in particular chapters 7 and 8 on “requirements of the risk management process” and “quality management system”)
190. El Mestari SZ, Lenzini G, Demirci H (2024) Preserving data privacy in machine learning systems. *Computers & Security*, Volume 137, 103605, ISSN 0167-4048. Online: <https://doi.org/10.1016/j.cose.2023.103605>
191. Elouataoui W et al. (2022) An advanced big data quality framework based on weighted metrics. *Big Data Cogn. Comput* 6(4), 153; Online: <https://doi.org/10.3390/bdcc6040153>
192. Elshamy R et al. (2023) Improving the efficiency of RMSProp optimizer by utilizing Nestrove in deep learning. *Sci Rep* 13, 8814. Online: <https://doi.org/10.1038/s41598-023-35663-x>
193. Emanuel EJ & Emanuel LL (1992) Four models of the physician-patient relationship, 267 *JAMA: The Journal of the American Medical Association* 2221–2226. Online: <https://jamanetwork.com/journals/jama/article-abstract/396718>
194. Emanuel EJ et al. (2000) What Makes Clinical Research Ethical? *JAMA*. 283(20):2701–2711. Online: doi:10.1001/jama.283.20.2701
195. Emanuel L et al., (2008) Advances in Patient Safety: New Directions and Alternative Approaches. Eds: Henriksen K et al. Rockville (MD) Agency for healthcare research and quality. Online: <https://www.ncbi.nlm.nih.gov/books/NBK43624/>
196. Emmert-Streib F et al. (2020) An Introductory Review of Deep Learning for Prediction Models With Big Data. *Front Artif Intell.* Feb 28;3:4. Online: doi: 10.3389/frai.2020.00004
197. EN ISO 14971:2019, Medical devices — Application of risk management to medical devices. Online: <https://www.iso.org/standard/72704.html>
198. Encyclopædia Britannica, entry "applied ethics": <https://www.britannica.com/topic/applied-ethics>

199. Encyclopædia Britannica, entry "metaethics": <https://www.britannica.com/topic/metaethics>
200. Encyclopædia Britannica, entry "normative ethics": <https://www.britannica.com/topic/normative-ethics>
201. Engen et al. (2016) Machine Agency in Human-Machine Networks; Impacts and Trust Implications. Online: arXiv:1602.08237. <https://doi.org/10.48550/arXiv.1602.08237>
202. Enid NH et al. (2009) Empirically understanding trust in medical technology. International Journal of Industrial Ergonomics, Volume 39, Issue 4, Pages 628-634. Online: <https://doi.org/10.1016/j.ergon.2009.01.004>
203. ENISA (2014) Technical Guideline on Security measures for Article 4 and Article 13a. Online: [https://resilience.enisa.europa.eu/article-13/guideline-on-security-measures-for-article-4-and-article-13a/TechnicalGuidelineonSecuritymeasuresforArticle4andArticle13a\\_version\\_1\\_0.pdf](https://resilience.enisa.europa.eu/article-13/guideline-on-security-measures-for-article-4-and-article-13a/TechnicalGuidelineonSecuritymeasuresforArticle4andArticle13a_version_1_0.pdf)
204. ENISA (2016) Definition of Cybersecurity - Gaps and overlaps in standardisation. Online: <https://www.enisa.europa.eu/publications/definition-of-cybersecurity>
205. ENISA (2022) Data Protection Engineering. Online: <https://www.enisa.europa.eu/publications/data-protection-engineering>
206. ENISA (2023) ENISA threat landscape: health sector. Online: <https://www.enisa.europa.eu/publications/health-threat-landscape>
207. Ettman CK & Galea S (2023) The Potential Influence of AI on Population Mental Health. JMIR Ment Health. Nov 16;10:e49936. Online: doi: 10.2196/49936.

### *European Commission*

208. European Commission (2018) European Group on Ethics in Science and New Technologies. Statement on artificial intelligence, robotics and 'autonomous' systems. Online: <https://op.europa.eu/en/publication-detail/-/publication/dfebe62e-4ce9-11e8-be1d-01aa75ed71a1> (Last accessed: 2024.08.21)
209. European Commission (2019) Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions on Building Trust in Human-Centric Artificial Intelligence, Brussels, 8.4.2019 COM(2019) 168 final. Online: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0064>
210. European Commission (2020a) Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics. COM/2020/64 final. Online: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0064>
211. European Commission (2020b) Website: Regulation on data governance – questions and answers. Online: [https://ec.europa.eu/commission/presscorner/detail/en/QANDA\\_20\\_2103](https://ec.europa.eu/commission/presscorner/detail/en/QANDA_20_2103) (Last accessed: 2024.09.19)
212. European Commission (2020c) Manufacturer Incident Report (MIR) form. Version 7.2.1.b. Online: <https://ec.europa.eu/docsroom/documents/41681>
213. European Commission (2020d) – Joint Research Centre, JRC AI Watch. Defining artificial intelligence. Online: doi: 10.2760/382730
214. European Commission (2021) Ethics by design and ethics of use approaches for artificial intelligence (Recommendations by a panel of experts at the request of the European Commission). Online: [https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence\\_he\\_en.pdf](https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence_he_en.pdf) (Last accessed 2024.09.01)
215. European Commission (2022a) Proposal for a DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive). Online: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0496>

216. European Commission (2022b) Questions & Answers: AI Liability Directive. Online: [https://ec.europa.eu/commission/presscorner/detail/en/QANDA\\_22\\_5793](https://ec.europa.eu/commission/presscorner/detail/en/QANDA_22_5793)
217. European Commission – Joint Research Centre (JRC) (2022c) Glossary on human-centric artificial intelligence. Online: <https://publications.jrc.ec.europa.eu/repository/handle/JRC129614>
218. European Commission (2023) The EU-U.S. “Terminology and Taxonomy for AI” containing 65 general AI terms in its first iteration. Online: <https://digital-strategy.ec.europa.eu/en/library/eu-us-terminology-and-taxonomy-artificial-intelligence>
219. European Commission (2024a) The official portal for European data: data.europa.eu. The Role of Artificial Intelligence in Processing and Generating New Data - An exploration of legal and policy challenges in open data ecosystems. Online: <https://data.europa.eu/sites/default/files/report/The%20Role%20of%20Artificial%20Intelligence%20in%20Processing%20and%20Generating%20New%20Data%20.pdf> (Last accessed: 2024.08.04)
220. European Commission (2024b): Website ‘Internal market, industry, entrepreneurship and SMEs’ – liability for defective products. Online: [https://single-market-economy.ec.europa.eu/single-market/goods/free-movement-sectors/liability-defective-products\\_en#:~:text=In%20the%20EU%20consumers%20can,interests%20of%20consumers%20and%20producers](https://single-market-economy.ec.europa.eu/single-market/goods/free-movement-sectors/liability-defective-products_en#:~:text=In%20the%20EU%20consumers%20can,interests%20of%20consumers%20and%20producers) (Last accessed: 2024.09.23)
221. European Commission (2024c) Europe’s choice. Political guidelines for the next European Commission 2024-2029. See page 9, security of health systems. Online: [https://commission.europa.eu/document/download/e6cd4328-673c-4e7a-8683-f63fb2cf648\\_en?filename=Political%20Guidelines%202024-2029\\_EN.pdf](https://commission.europa.eu/document/download/e6cd4328-673c-4e7a-8683-f63fb2cf648_en?filename=Political%20Guidelines%202024-2029_EN.pdf)
222. European Commission (2024d) Website on health technology assessment. Online: [https://health.ec.europa.eu/health-technology-assessment/overview\\_en](https://health.ec.europa.eu/health-technology-assessment/overview_en) (Last accessed: 2024.08.20).
223. European Commission (2024e) The Role of Artificial Intelligence in Processing and Generating New Data - An exploration of legal and policy challenges in open data ecosystems. Online: <https://data.europa.eu/sites/default/files/report/The%20Role%20of%20Artificial%20Intelligence%20in%20Processing%20and%20Generating%20New%20Data%20.pdf> (Last accessed: 2024.08.04)
224. European Commission (2025a) European action plan on the cybersecurity of hospitals and healthcare providers. Online: <https://digital-strategy.ec.europa.eu/en/library/european-action-plan-cybersecurity-hospitals-and-healthcare-providers>
225. European Commission (2025b) Data Act explained. Online: <https://digital-strategy.ec.europa.eu/en/factpages/data-act-explained>
226. European Commission (2025c) Website: Personalised medicine. Online: [https://research-and-innovation.ec.europa.eu/research-area/health/personalised-medicine\\_en](https://research-and-innovation.ec.europa.eu/research-area/health/personalised-medicine_en)
227. European Commission (2025d) AI literacy – questions & answers. Online: <https://digital-strategy.ec.europa.eu/en/faqs/ai-literacy-questions-answers>

## *European Union (EU)*

228. European Union (2000) Charter of fundamental rights of the European Union. Online: [https://www.europarl.europa.eu/charter/pdf/text\\_en.pdf](https://www.europarl.europa.eu/charter/pdf/text_en.pdf)
229. European Union (2007) Treaty on the Functioning of the European Union of 13 December 2007 – consolidated version (OJ C 202, 7.6.2016, pp. 47–360). Online: <https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:12012E/TXT:en:PDF>
230. EU (2025) European Council. Council conclusions on personalised medicine for patients. Online: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ%3AC%3A2015%3A421%3AFULL>
231. EU (2016a) Consolidated version of the treaty on European Union. Online: [https://eur-lex.europa.eu/resource.html?uri=cellar:2bf140bf-a3f8-4ab2-b506-fd71826e6da6.0023.02/DOC\\_1&format=PDF](https://eur-lex.europa.eu/resource.html?uri=cellar:2bf140bf-a3f8-4ab2-b506-fd71826e6da6.0023.02/DOC_1&format=PDF)

- 232. EU (2016b) Regulation (EU) 2016/679. General Data Protection Regulation (GDPR). Online: <https://eur-lex.europa.eu/eli/reg/2016/679/oj>
- 233. EU (2017a) Regulation (EU) 2017/745 on medical devices. Online: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32017R0745>
- 234. EU (2017b) Regulation (EU) 2017/746 on in vitro diagnostic medical devices. "In vitro diagnostics medical devices Regulation" (IVDR). Online: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32017R0746>
- 235. EU (2020) EU horizon 2020 project: PANACEA - Protection and privAcy of hospital and health iNfrastructure with smArt Cyber sEcurity and cyber threat toolkit for dAta and people. Online: <https://www.panacearesearch.eu/> See the "results" section for publications and outputs of this H2020 project.
- 236. EU (2022) Data Act. Online: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2022%3A68%3AFIN>
- 237. EU (2024a) Regulation laying down harmonised rules on artificial intelligence ("AI Act"). Online: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689>. For a version that can be easily navigated, see <https://artificialintelligenceact.eu/the-act/>
- 238. EU (2024b) The digital services act. Website: [https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act\\_en](https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en)
- 239. EU (2024c) Europe's choice. Political guidelines for the next European Commission 2024-2029. See page 9, security of health systems. Online: [https://commission.europa.eu/document/download/e6cd4328-673c-4e7a-8683-f63ffb2cf648\\_en](https://commission.europa.eu/document/download/e6cd4328-673c-4e7a-8683-f63ffb2cf648_en)

### *EU HLEG (EU high-level expert group)*

- 240. EU HLEG (2019) Independent high-level expert group on artificial intelligence set up by the European Commission: Ethics guideline for trustworthy AI. Online: [https://ec.europa.eu/futurium/en/ai-alliance-consultation\\_1.html](https://ec.europa.eu/futurium/en/ai-alliance-consultation_1.html)
- 241. EU HLEG (2020) Independent high-level expert group on artificial intelligence set up by the European Commission: Assessment list for trustworthy AI (ALTAI). Online: <https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altais-self-assessment>

### *EU MDCG - medical devices coordination group*

- 242. EU Medical Devices Coordination Group, MDCG (2019) MDCG 2019-11: Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745 – MDR and Regulation (EU) 2017/746 – IVDR. Online: [MDCG 2019-11](#).
- 243. EU Medical Devices Coordination Group (MDCG) (2020a) MDCG 2020-1: Guidance on Clinical Evaluation (MDR) / Performance Evaluation (IVDR) of Medical Device Software (MDSW). Online: [MDCG 2020-1](#)
- 244. EU Medical Devices Coordination Group (MDCG) (2023) MDCG 2023-4: Medical Device Software (MDSW) – Hardware combinations. *Guidance on MDSW intended to work in combination with hardware or hardware components.* Online: [https://health.ec.europa.eu/document/download/b2c4e715-f2b4-4d24-af60-056b5d41a72e\\_en?filename=md\\_mdcg\\_2023-4\\_software\\_en.pdf](https://health.ec.europa.eu/document/download/b2c4e715-f2b4-4d24-af60-056b5d41a72e_en?filename=md_mdcg_2023-4_software_en.pdf)
- 245. EU Medical Devices Coordination Group (MDCG) (2024a) MDCG 2024-3: Guidance on content of the Clinical Investigation Plan for clinical investigations of medical devices. Online: [MDCG 2024-3](#)
- 246. EU Medical Devices Coordination Group (MDCG) (2024b) MDCG 2024-5: Guidance on content of the Investigator's Brochure for clinical investigations of medical devices. Online: [https://health.ec.europa.eu/document/download/ee7ee8eb-841a-4085-a8dc-af6d55ebf1bd\\_en?filename=mdcg\\_2024-5\\_en.pdf](https://health.ec.europa.eu/document/download/ee7ee8eb-841a-4085-a8dc-af6d55ebf1bd_en?filename=mdcg_2024-5_en.pdf)

- 247. EU Medical Devices Coordination Group (MDCG) (2025) MDCG 2025-6: Interplay between the Medical Regulation (MDR) & In vitro Diagnostic Medical Devices Regulation (IVDR) and the Artificial Intelligence Act (AIA). Online: [https://health.ec.europa.eu/latest-updates/mdcg-2025-6-faq-interplay-between-medical-devices-regulation-vitro-diagnostic-medical-devices-2025-06-19\\_en](https://health.ec.europa.eu/latest-updates/mdcg-2025-6-faq-interplay-between-medical-devices-regulation-vitro-diagnostic-medical-devices-2025-06-19_en)
- 248. European Academy of Paediatrics (2023) Shortages of medical devices for children. Online: <https://www.eapaediatrics.eu/wp-content/uploads/2023/08/EAP-Press-Release-Shortages-of-Medical-Devices-for-Children-18-August-2023.pdf> (Last accessed: 2024.09.07)

## *European Parliament*

- 249. European Parliament (2020) European Parliament resolution of 20 October 2020 with recommendations to the Commission on a civil liability regime for artificial intelligence (2020/2014(INL)). Online: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020IP0276>
- 250. European Parliament (2022a) New product liability directive. In 'A Europe fit for the digital age'. Online: <https://www.europarl.europa.eu/legislative-train/theme-a-europe-fit-for-the-digital-age/file-new-product-liability-directive> (Last accessed: 2024.09.23)
- 251. European Parliament (2022b) Scientific Foresight Unit (STOA) Artificial intelligence in healthcare. Applications, risks and ethical and societal impacts. Online: [https://www.europarl.europa.eu/thinktank/en/document/EPRS\\_STU\(2022\)729512](https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2022)729512)
- 252. European Parliament (2025) – legislative train schedule: AI liability Directive. Online: <https://www.europarl.europa.eu/legislative-train/theme-a-europe-fit-for-the-digital-age/file-ai-liability-directive?p3373>
- 253. EQUATOR network (2025) Enhancing the QUAlity and Transparency Of health Research. Online: <https://www.equator-network.org/> (Last accessed 2024.08.06)

## **F**

- 254. Fadlallah R et al. (2024) The effects of public health and social measures (PHSM) implemented during the COVID-19 pandemic: An overview of systematic reviews. Cochrane evidence synthesis and methods. Cochrane Ev Synth. 2024; 2:e12055. Online: doi:10.1002/cesm.12055
- 255. Faes L et al. (2022) Artificial Intelligence and Statistics: Just the Old Wine in New Wineskins? Front Digit Health. Jan 26;4:833912. Online: doi: 10.3389/fdgth.2022.833912
- 256. Fagan F (2025) Reducing proxy discrimination. arXiv:2501.03946. Online: <https://doi.org/10.48550/arXiv.2501.03946>
- 257. Faisal R (2020) What are model parameters in deep learning. Medium website. Online: <https://medium.com/analytics-vidhya/what-are-model-parameters-in-deep-learning-and-how-to-calculate-it-de96476caab>
- 258. Falkowitz R (2015) On value streams, chains and life cycles. Online: [https://www.3cs.ch/on\\_value\\_streams\\_chains\\_and\\_life\\_cycles/](https://www.3cs.ch/on_value_streams_chains_and_life_cycles/) (Last accessed: 2025.01.08)
- 259. Farah L et al. (2023) Are current clinical studies on artificial intelligence-based medical devices comprehensive enough to support a full health technology assessment? A systematic review. Artif Intell Med. 2023 Jun;140:102547. Online: doi: 10.1016/j.artmed.2023.102547.
- 260. FarnamStreet (2024) John Stuart Mill's Philosophy of Equality. Online: <https://fs.blog/john-stuart-mills-equality/> (Last accessed 2024.09.07)
- 261. Feldman J (2001) Artificial intelligence. In: 'Artificial intelligence in cognitive science' in: The International Encyclopedia of the Social & Behavioural Sciences, Vol 2, pp. 792-796.
- 262. Feldman M et al. (2014) Certifying and removing disparate impact. arXiv:1412.3756. Online: <https://doi.org/10.48550/arXiv.1412.3756>

263. Fensel D (2001) Ontologies. In: *Ontologies*. Springer, Berlin, Heidelberg. Online: [https://doi.org/10.1007/978-3-662-04396-7\\_2](https://doi.org/10.1007/978-3-662-04396-7_2)
264. Ferreiro Y (2023) What is Continuous Learning? Revolutionizing Machine Learning & Adaptability. Online: <https://www.datacamp.com/blog/what-is-continuous-learning>
265. Ferrero L ed. (2022) The Routledge Handbook of Philosophy of Agency. Chapter: An introduction to the philosophy of agency. Online: <https://philarchive.org/archive/FERAIT-7> (Last accessed 13.5.2024).
266. Feuerriegel S et al. (2024) Causal machine learning for predicting treatment outcomes. *Nat Med* 30, 958–968. Online: <https://doi.org/10.1038/s41591-024-02902-1>
267. Finlayson SG et al. (2021) The Clinician and Dataset Shift in Artificial Intelligence. *N Engl J Med.* 2021 Jul 15;385(3):283–286. Online: doi: 10.1056/NEJMc2104626.
268. Firt E (2024) Addressing corrigibility in near-future AI systems. *AI Ethics*. Online: <https://doi.org/10.1007/s43681-024-00484-9> (Last accessed: 2024.08.12)
269. Fisher E et al. (2023) Occupational Safety and Health Equity Impacts of Artificial Intelligence: A Scoping Review. *Int J Environ Res Public Health.* Jun 24;20(13):6221. Online: doi: 10.3390/ijerph20136221
270. FitzGerald C & Hurst S (2027). Implicit bias in healthcare professionals: a systematic review. *BMC Med Ethics.* Mar 1;18(1):19. Online: doi: 10.1186/s12910-017-0179-8.
271. Fjeld J (2020) Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI (January 15, 2020). Berkman Klein Center for Internet & Society at Harvard University. Research Publication No. 2020-1, Available at SSRN: <https://ssrn.com/abstract=3518482> or <http://dx.doi.org/10.2139/ssrn.3518482>
272. Fleetwood M (2023) Gödel's Incompleteness Theorem and the Limits of AI. Medium.com.
273. Floridi et al. (2018) AI4people - an ethical framework for a good AI society: opportunities, risks, principles and recommendations. *Minds and Machines,* 28(4):689–707. Online: <https://doi.org/10.1007/s11023-018-9482-5>
274. Floridi L & Cowls J (2019) A unified framework of five principles for AI in society. *Harvard Data Science Review.* 1.1 Online: <https://doi.org/10.1162/99608f92.8cd550d1>
275. Floridi L et al. (2019) Establishing the rules for building trustworthy AI. *Nat Mach Intell* 1, 261–262. <https://doi.org/10.1038/s42256-019-0055-y>.
276. Frasca M et al. (2024) Explainable and interpretable artificial intelligence in medicine: a systematic bibliometric review. *Discov Artif Intell* 4, 15. Online: <https://doi.org/10.1007/s44163-024-00114-7>
277. Fraser H et al. (2024) Acceptable Risks in Europe's Proposed AI Act: Reasonableness and Other Principles for Deciding How Much Risk Management Is Enough. *European Journal of Risk Regulation.* 15(2):431–446. Online: doi:10.1017/err.2023.57
278. French Government (2025) Statement on Inclusive and Sustainable Artificial Intelligence for People and the Planet. Online: <https://www.elysee.fr/en/emmanuel-macron/2025/02/11/state-ment-on-inclusive-and-sustainable-artificial-intelligence-for-people-and-the-planet>
279. Frey CB & Osborne MA (2017) The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, Volume 114, pages 254-280. Online: <https://doi.org/10.1016/j.techfore.2016.08.019> .
280. FUTURE-AI consortium (2023) Best practices for trustworthy AI in medicine. Future AI guidelines: traceability. <https://future-ai.eu/principle/traceability/>
281. FUTURE-AI consortium (2025): international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ.* 2025 Feb 17;388:r340. doi: 10.1136/bmj.r340. Erratum for: *BMJ.* 2025 Feb 5;388:e081554. doi: 10.1136/bmj-2024-081554.
282. Future of life institute (2024a) Asolimar AI principles. Online: <https://futureoflife.org/open-letter/ai-principles/>

283. Future of life institute (2024b) Predictable AI. Online: <https://www.predictable-ai.org/> (Last accessed: 2024.08.16)

## G

284. Gadd CS et al. (2016) Creating advanced health informatics certification. *J Am Med Inform Assoc.* 2016 Jul;23(4):848-50. doi: 10.1093/jamia/ocw089
285. Ganapathi S et al. (2022) Tackling bias in AI health datasets through the STANDING Together initiative. *Nat Med.* Nov;28(11):2232-2233. Online: doi: 10.1038/s41591-022-01987-w
286. Garrett BL & Rudin C (2023) Interpretable algorithmic forensics. *Proc Natl Acad Sci U S A.* 2023 Oct 10;120(41):e2301842120. Online: doi: 10.1073/pnas.2301842120. Epub 2023 Oct 2.
287. Gazquez-Garcia J, Sánchez-Bocanegra C, Sevillano J (2025) AI in the Health Sector: Systematic Review of Key Skills for Future Health Professionals. *JMIR Med Educ* 11:e58161. Online: <https://mededu.jmir.org/2025/1/e58161>; DOI: 10.2196/58161
288. GDPR.EU (2025) Data Protection Impact Assessment (DPIA). Online: <https://gdpr.eu/data-protection-impact-assessment-template/>
289. Gebru T et al. (2021) Datasheets for datasets. *arXiv:1803.09010*. Online: <https://doi.org/10.48550/arXiv.1803.09010> (Last accessed: 2024.08.20).
290. Geirhos, R et al. (2020) Shortcut learning in deep neural networks. *Nat Mach Intell* 2, 665–673. Online: <https://doi.org/10.1038/s42256-020-00257-z>
291. Geis JR et al. (2019) Ethics of Artificial Intelligence in Radiology: Summary of the Joint European and North American Multisociety Statement. *J Am Coll Radiol.* Nov;16(11):1516-1521. doi: 10.1016/j.jacr.2019.07.028
292. George Washington University – Milken Institute School of Public Health (2024). Equity vs equality: what's the difference? Online: <https://onlinepublichealth.gwu.edu/resources/equity-vs-equality/#:~:text=Equality%20means%20each%20individual%20or,to%20reach%20an%20equal%20outcome>
293. Gershgorn G (2018) If AI Is Going to Be the World's Doctor, It Needs Better Textbooks, Quartz, September 6 Online: <https://qz.com/1367177/if-ai-is-going-to-be-the-worlds-doctor-it-needs-better-textbooks> (Last accessed: 2024.09.07)
294. Ghassemi M et al. (2021) The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health.* Nov;3(11):e745-e750. doi: 10.1016/S2589-7500(21)00208-9
295. Gibson C (2024) How AI Is Transforming Cyber Threats in 2025. Online: <https://builtin.com/artificial-intelligence/ai-transforming-cyber-threats>
296. Giles-Clark HJ et al. (2023) Should we use composite outcomes in obstetric clinical prediction models? *Eur J Obstet Gynecol Reprod Biol.* Jun;285:193-197. Online: doi: 10.1016/j.ejogrb.2023.04.031.
297. Gill AS (2021) Aligning AI governance globally: lessons from current practice. Stiftung Entwicklung und Frieden (sef) – INEF. Online: [https://static1.squarespace.com/static/651acb4c077bed428243c484/t/6532a53734300834518eb884/1697817912176/GT-A\\_2021-03\\_en.pdf](https://static1.squarespace.com/static/651acb4c077bed428243c484/t/6532a53734300834518eb884/1697817912176/GT-A_2021-03_en.pdf)
298. Giordano D (2020) Towards data science website. 7 tips to choose the best optimizer. Online: <https://towardsdatascience.com/7-tips-to-choose-the-best-optimizer-47bb9c1219e/>
299. Gilpin LH et al. (2018). Explaining explanations: An overview of interpretability of machine learning. In 2018 IEEE 5th international conference on data science and advanced analytics (DSAA) (pp. 80–89). arXiv:1806.00069
300. Github (2025) The LLM evaluation guidebook. Online: <https://github.com/huggingface/evaluation-guidebook>

301. Giuffrè M & Shung DL (2023) Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *npj Digit. Med.* 6, 186. Online: <https://doi.org/10.1038/s41746-023-00927-3>
302. Global Harmonization Task Force (now IMDRF) (2005) GHTF/SG1/N41R9:2005. Essential Principles of Safety and Performance of Medical Devices. Online: <https://www.imdrf.org/sites/default/files/docs/ghtf/final/sg1/technical-docs/ghtf-sg1-n41r9-2008-principles-safety-performance-050520.pdf>
303. Global harmonisation task force (now IMDRF) (2006) GHTF/SG2/ N54R8:2006. Medical devices post market surveillance: global guidance for adverse event reporting of medical devices. Online: <https://www.imdrf.org/sites/default/files/docs/ghtf/final/sg2/technical-docs/ghtf-sg2-n54r8-guidance-adverse-events-061130.pdf>
304. Global Harmonization Task Force (now IMDRF) (2011) GHTF/SG1/N70:2011: Label and Instructions for Use for Medical Devices. Online: <https://www.imdrf.org/sites/default/files/docs/ghtf/archived/sg1/technical-docs/ghtf-sg1-n70-2011-label-instruction-use-medical-devices-110916.pdf>
305. Global Harmonization Task Force (now IMDRF) (2012) GHTF/SG5/N7:2012 Clinical Evidence for IVD medical devices – Scientific Validity Determination and Performance Evaluation. Online: <https://www.imdrf.org/sites/default/files/docs/ghtf/final/sg5/technical-docs/ghtf-sg5-n7-2012-scientific-validity-determination-evaluation-121102.pdf>
306. Glover E (2024) AI in Cybersecurity: The Good and the Bad. Online: <https://builtin.com/articles/ai-and-cybersecurity>
307. Glymour CN (2001) *The mind's arrows: Bayes nets and graphical causal models in psychology*. MIT press. Online: <https://doi.org/10.7551/mitpress/4638.001.0001>
308. Goertzel B (2014) Artificial general intelligence: concept, state of the art, and future prospects. *Journal of Artificial Intelligence* 5:1-48. Online: <https://doi.org/10.2478/jagi-2014-0001>
309. Goldsack JC (2020) Verification, analytical validation, and clinical validation (V3): the foundation of determining fit-for-purpose for Biometric Monitoring Technologies (BioMeTs). *NPJ Digit Med.* 2020 Apr 14;3:55. doi: 10.1038/s41746-020-0260-4. PMID: 32337371; PMCID: PMC7156507.
310. Goodman B & Flaxman S (2016, 2017) European Union regulations on algorithmic decision-making and a "right to explanation". arXiv:1606.08813. <https://doi.org/10.1609/aimag.v38i3.2741>
311. Google Machine Learning glossary. Online: <https://developers.google.com/machine-learning/glossary?hl=en>
312. Gonsalves T & Upadhyay J (2021) Chapter Eight - Integrated deep learning for self-driving robotic cars, Editor(s): Rabindra Nath Shaw, Ankush Ghosh, Valentina E. Balas, Monica Bianchini, Artificial Intelligence for Future Generation Robotics, Elsevier, Pages 93-118. <https://doi.org/10.1016/B978-0-323-85498-6.00010-1>
313. Gopnik A (2012) Scientific thinking in young children: Theoretical advances, empirical research, and policy implications. *Science*, 337(6102), 1623–1627. Online: <https://doi.org/10.1126/science.1223416>
314. Gopnik A (2024) Causal learning. In: Open encyclopedia of cognitive science. Online: <https://oecs.mit.edu/pub/i1om74mo/release/1>
315. Governance.ai (2024) Online: <https://www.governance.ai/>
316. Gray C (2023) More than Extraction: Rethinking Data's Colonial Political Economy, *International Political Sociology*, Volume 17, Issue 2, June 2023, olad007, <https://doi.org/10.1093/ips/olad007>
317. Griesinger CB et al. (2016) Validation of Alternative In Vitro Methods to Animal Testing: Concepts, Challenges, Processes and Tools. *Adv Exp Med Biol.* 856:65-132. Online: doi: 10.1007/978-3-319-33826-2\_4.

318. Griesinger CB et al. (2022) chapter „medicine and healthcare” in: European Commission, Joint Research Centre, Balahur, A., Jenet, A., Hupont Torres, I. et al. Data quality requirements for inclusive, non-biased and trustworthy AI – Putting science into standards, Publications Office of the European Union, 2022. Online: <https://data.europa.eu/doi/10.2760/365479>
319. Griffiths TL et al. (2010) Probabilistic models of cognition: exploring representations and inductive biases. *Trends Cogn Sci.* Aug;14(8):357-64. doi: 10.1016/j.tics.2010.05.004
320. Grote T & Berens P (2019) On the ethics of algorithmic decision-making in healthcare. *Journal of medical ethics.* 46:205-211. Online: <https://jme.bmjjournals.org/content/46/3/205.long>
321. Grote T & Keeling G (2022) On Algorithmic Fairness in Medical Practice. *Camb Q Healthc Ethics.* Jan;31(1):83-94. Online: doi: 10.1017/S0963180121000839.
322. Guerra-Farfan E et al. (2023) Clinical practice guidelines: The good, the bad, and the ugly. *Injury.* 2023 May;54 Suppl 3:S26-S29. doi: 10.1016/j.injury.2022.01.047
323. Guni A et al., (2024) Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies Using AI (QUADAS-AI): Protocol for a Qualitative Study. *JMIR Res Protoc.* 2024 Sep 18;13:e58202. doi: 10.2196/58202
324. Gunning D & Aha D (2019) DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine,* 40(2), 44-58. <https://doi.org/10.1609/aimag.v40i2.2850>
325. Gunning, D et al. (2021) DARPA's explainable AI (XAI) program: A retrospective. *Applied AI Letters,* 2: e61. <https://doi.org/10.1002/ail.2.61>

## H

326. Haeri MA & Zweig KA (2020) The Crucial Role of Sensitive Attributes in Fair Classification. 2020 IEEE Symposium Series on Computational Intelligence (SSCI), Canberra, ACT, Australia, pp. 2993-3002, doi: 10.1109/SSCI47803.2020.9308585.
327. Hagendorff T (2020) The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds & Machines* 30, 99–120. <https://doi.org/10.1007/s11023-020-09517-8>
328. Hájek A (2001) Counterfactual Reasoning, Quantitative: Philosophical Aspects. Editor(s): Neil J. Smelser, Paul B. Baltes. International Encyclopedia of the Social & Behavioral Sciences. Pergamon, p. 2872-2874. ISBN 9780080430768. Online: <https://doi.org/10.1016/B0-08-043076-7/01015-9>
329. Halpern JY & Pearl J (2005a) Causes and explanations: A structural-model approach. Part I: Causes. *British Journal for the Philosophy of Science* 56 (4):843-887. Online: 10.1093/bjps/axi147. Available also on arXiv. Online: <https://doi.org/10.48550/arXiv.1301.2275>
330. Halpern JY & Pearl J (2005b). Causes and Explanations: A Structural-Model Approach. Part II: Explanations. *British Journal for the Philosophy of Science* 56 (4):889-911. Online: 10.1093/bjps/axi148. Available also on arXiv. Online: <https://doi.org/10.48550/arXiv.cs/0208034>
331. Halpern JY (2015) A Modification of the Halpern-Pearl Definition of Causality. arXiv. Online: <https://doi.org/10.48550/arXiv.1505.00162>
332. Hancock PA (2023) How and why humans trust: A meta-analysis and elaborated model. *Front Psychol.* Mar 27; 14:1081086. Online: doi: 10.3389/fpsyg.2023.1081086
333. Hansen M et al. (2016) “Protection Goals for Privacy Engineering” in 2015 IEEE Security and Privacy Workshops (SPW), 2015, San Jose, CA: IEEE, pp. 159-166. [44] Office of the Federal Chief Information Officer, CIRCULAR No. A-130. Online: <https://ieeexplore.ieee.org/document/7163220>
334. Hansson SO (2020) John Stuart Mill and the Conflicts of Equality. *J Ethics* 26, 433–453. Online: <https://doi.org/10.1007/s10892-022-09393-7>
335. Hardt M et al. (2016) Equality of Opportunity in Supervised Learning. arXiv:1610.02413. Online: <https://doi.org/10.48550/arXiv.1610.02413>

336. Harkin KR et al. (2024) Lifecycle evaluation of medical devices: supporting or jeopardizing patient outcomes? A comparative analysis of evaluation models. *Int J Technol Assess Health Care.* 2024 Jan;5;40(1):e2. doi: 10.1017/S026646232300274X
337. Harman GH (1965) The inference to the best explanation. *The Philosophical Review.* Vol. 74, No. 1 (Jan., 1965), pp. 88-95. Pdf version of scanned paper available online from Carnegie Mellon University: <https://www.andrew.cmu.edu/user/kk3n/philsciclass/harman.pdf> (last accessed: 27.4.2025).
338. HealthAI – The global agency for responsible AI in health (2024) Mapping AI governance in health. From global regulatory alignments to LMICs' Policy Developments. Online: [https://static1.squarespace.com/static/651acb4c077bed428243c484/t/675183aa252e7e556dbba30a/1733395380191/HealthAI\\_Global+Landscape+Report\\_Oct.2024.pdf](https://static1.squarespace.com/static/651acb4c077bed428243c484/t/675183aa252e7e556dbba30a/1733395380191/HealthAI_Global+Landscape+Report_Oct.2024.pdf)
339. Healthcare-in-europe.com (2024) Our future health: shifting from curative to preventive care. Online: <https://healthcare-in-europe.com/en/news/our-future-health-curative-preventive-care.html>
340. Hebb D.O. (1949) The organisation of behaviour – a neuropsychological theory. Wiley, New York. Online accessible via: [https://pure.mpg.de/pubman/item/item\\_2346268\\_3/component/file\\_2346267/Hebb\\_1949\\_The\\_Organisation\\_of\\_Behavior.pdf](https://pure.mpg.de/pubman/item/item_2346268_3/component/file_2346267/Hebb_1949_The_Organisation_of_Behavior.pdf)
341. Heeger DJ et al. (1996) Computational models of cortical visual processing. *Proc Natl Acad Sci U S A.* Jan 23;93(2):623-7. Online: doi: 10.1073/pnas.93.2.623.
342. Hempel CG & Oppenheim P (1948) Studies in the logic of explanation. *Philos. Sci.*, 15 (2), pp. 135-175. Online: <https://www.sfu.ca/~jillmc/Hempel%20and%20Oppenheim.pdf>
343. Hermann K et al. (2024) On the foundations of shortcut learning. arXiv. Online: arXiv:2310.16228
344. Herzog C (2022) On the risk of confusing interpretability with explicability. *AI Ethics* 2, 219–225 (2022). <https://doi.org/10.1007/s43681-021-00121-9>
345. Hewitt-Taylor J (2004) Clinical guidelines and care protocols. *Intensive Crit Care Nurs.* 2004 Feb;20(1):45-52. doi: 10.1016/j.iccn.2003.08.002. PMID: 14726253.
346. Hicks SA et al. (2022) On evaluation metrics for medical applications of artificial intelligence. *Sci Rep.* 2022 Apr 8;12(1):5979. doi: 10.1038/s41598-022-09954-8.
347. Hinton G et al. (2018) Coursera neural networks for machine learning lecture 6. Online: [https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture\\_slides\\_lec6.pdf](https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf)
348. Hjeij M & Vilks A (2023) A brief history of heuristics: how did research on heuristics evolve? *Humanit Soc Sci Commun* 10, 64 (2023). <https://doi.org/10.1057/s41599-023-01542-z>
349. Höffe O et al. (eds.) (1986) Lexikon der Ethik. C.H. Beck, München.
350. Hoffman S (2021) The Emerging Hazard of AI-Related Health Care Discrimination. *Hastings Cent Rep.* 2021 Jan;51(1):8-9. doi: 10.1002/hast.1203. Epub 2020 Dec 14. PMID: 33315263
351. Holzinger A (2021) Explainable AI and Multi-Modal Causability in Medicine. *I Com (Berl).* Jan 26;19(3):171-179. Online: doi: 10.1515/icom-2020-0024. Epub 2021 Jan 15.
352. Holzinger A et al. (2019) Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip Rev Data Min Knowl Discov.* Jul-Aug;9(4):e1312. Online: doi: 10.1002/widm.1312. Epub 2019 Apr 2.
353. Holzinger A et al. (2022) Explainable AI Methods - A Brief Overview. In: Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, KR., Samek, W. (eds) xxAI – Beyond Explainable AI. xxAI 2020. Lecture Notes in Computer Science, vol 13200. Springer, Cham. [https://doi.org/10.1007/978-3-031-04083-2\\_2](https://doi.org/10.1007/978-3-031-04083-2_2)
354. Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Acad Sci U S A.* Apr;79(8):2554-8. Online: doi: 10.1073/pnas.79.8.2554.

355. Hoque S et al., (2023) Deep Learning model integrity checking mechanism using watermarking technique. arXiv:2301.12333. <https://doi.org/10.48550/arXiv.2301.12333>
356. Hosseinzadeh Taher MR et al. (2021) A Systematic Benchmarking Analysis of Transfer Learning for Medical Image Analysis. *Domain Adapt Represent Transf Afford Healthc AI Resour Divers Glob Health* (2021). 2021 Sep-Oct;12968:3-13. Online: doi: 10.1007/978-3-030-87722-4\_1.
357. Howard FM et al. (2021) The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nat Commun* 12, 4423 (2021). Online: <https://doi.org/10.1038/s41467-021-24698-1>
358. Howell MD (2024) Generative artificial intelligence, patient safety and healthcare quality: a review. *BMJ Qual Saf*. Oct 18;33(11):748-754. Online: doi: 10.1136/bmjqqs-2023-016690
359. Huang J (2020) RMSprop. Cornell University website. Online: <https://optimisation.cbe.cornell.edu/index.php?title=RMSProp>
360. Huang H (2023) Performance of ChatGPT on Registered Nurse License Exam in Taiwan: A Descriptive Study. *Healthcare (Basel)*. Oct 30;11(21):2855. doi: 10.3390/healthcare11212855
361. Huang J et al. (2024a) Evaluating large language model (LLM) systems: metrics, challenges and best practices. Online: Evaluating Large Language Model (LLM) systems: Metrics, challenges, and best practices | by Jane Huang | Data Science at Microsoft | Medium
362. Huang S et al. (2024b) AI Technology panic - is AI Dependence Bad for Mental Health? A Cross-Lagged Panel Model and the Mediating Roles of Motivations for AI Use Among Adolescents. *Psychol Res Behav Manag*. Mar 12;17:1087-1102. Online: doi: 10.2147/PRBM.S440889
363. HuggingFace (2025) AI energy scoreboard. Online: <https://huggingface.co/AIEnergyScore>
364. Hyzy M (2024) Why we need AI governance now. Article on 'built in' website. Online: <https://builtin.com/articles/we-need-ai-governance>

365. IBM (2024) AI Fairness 360. Online: <https://aif360.res.ibm.com/> (Last accessed: 2024.08.07)
366. IBM (2024) Shedding light on AI bias with real world examples. Online: <https://www.ibm.com/think/topics/shedding-light-on-ai-bias-with-real-world-examples>
367. IBM (2024) What is AI governance? Online: <https://www.ibm.com/topics/ai-governance>
368. IBM (2024) What is automation? Online: <https://www.ibm.com/topics/automation>
369. IBM (2024) What is gradient descent? Online: <https://www.ibm.com/think/topics/gradient-descent>
370. Ibrahim MH et al. (2021) Detecting Important Patterns Using Conceptual Relevance Interestingness Measure. Online: arXiv:2110.11262. <https://doi.org/10.48550/arXiv.2110.11262>
371. IEC (2015) IEC62366-1:2015. Medical devices – Application of usability engineering to medical devices. Online: <https://www.iso.org/standard/63179.html>
372. IEEE (2016) IEEE Standard for System, Software, and Hardware Verification and Validation. Online: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8055462> (paywall)
373. IEEE (2017) Ethically aligned design. A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems. Version 2 - for public discussion. Online: <https://standards.ieee.org/industry-connections/ec/ead-v1/> and [https://www.academia.edu/38746211/ETHICALLY\\_ALIGNED DESIGN\\_A\\_Vision\\_for\\_Prioritizing\\_Human\\_Wellbeing\\_with\\_Artificial\\_Intelligence\\_and\\_Autonomous\\_Systems](https://www.academia.edu/38746211/ETHICALLY_ALIGNED DESIGN_A_Vision_for_Prioritizing_Human_Wellbeing_with_Artificial_Intelligence_and_Autonomous_Systems)
374. INAHTA (2020) Announcing the new definition of HTA! (website). Online: <https://www.inahta.org/2020/05/announcing-the-new-definition-of-hta/>

375. Infosec Institute (2023) What does a system administrator do and how to become one? (3<sup>rd</sup> December 2023) Online: <https://www.infosecinstitute.com/resources/professional-development/what-does-a-system-administrator-do-and-how-to-become-one/>
376. Institute for the future of work (2024) Algorithmic impact assessment - knowledge hub. Online: <https://www.ifow.org/knowledge-hub-themes/algorithmic-impact-assessment> (Last accessed 13.5.2024)
377. Institute of Medicine (2011) Committee on Standards for Developing Trustworthy Clinical Practice Guidelines (2011) Clinical Practice Guidelines We Can Trust. Graham R, Mancher M, Miller Wolman D, Greenfield S, Steinberg E, editors. Washington (DC), USA: National Academies Press. Online: <https://www.ncbi.nlm.nih.gov/books/NBK209539/>
378. International Association for Impact Assessment (2024) Impact assessment. Online: <https://www.iaia.org/wiki-details.php?ID=4> (Last accessed 13.5.2024)

## IMDRF

379. International Medical Device Regulators Forum (IMDRF) (2006-2012 GHTF Study Group 2 - Post-market Surveillance/Vigilance. Online: <https://www.imdrf.org/documents/ghtf-final-documents/ghtf-study-group-2-post-market-surveillancevigilance>
380. International Medical Devices Regulators Forum (IMDRF) (2013) IMDRF/SaMD WG/N10FINAL:2013. Software as a medical device (SaMD) - key definitions. Online: <https://www.imdrf.org/working-groups/software-medical-device-samd>
381. International Medical Device Regulators Forum (IMDRF) (2015) IMDRF/SaMD WG/N23FINAL:2015 Software as a Medical Device (SaMD): Application of Quality Management System. Online: [IMDRF - SAMD and application of QMS \(N23:2015\)](#)
382. International Medical Device Regulators Forum (IMDRF) (2017a) IMDRF/SaMD WG/N41FINAL:2017. Software as a medical device (SaMD): clinical evaluation. Online: [https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-170921-samdn41-clinical-evaluation\\_1.pdf](https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-170921-samdn41-clinical-evaluation_1.pdf)
383. International Medical Device Regulators Forum (IMDRF) (2017b) IMDRF/GRRP WG/N47 FINAL:2018 IMDRF Good Regulatory Review Practices Group. Online: <https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-181031-grrp-essential-principles-n47.pdf>
384. International Medical Device Regulators Forum (IMDRF) (2018, 2024 edition 2) Essential Principles of Safety and Performance of Medical Devices and IVD Medical Devices. Online: <https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-181031-grrp-essential-principles-n47.pdf> and (edition 2) <https://www.imdrf.org/sites/default/files/2024-04/IMDRF%20GRRP%20WG%20N47%20%28Edition%202%29.pdf>
385. International Medical Device Regulators Forum (IMDRF) (2019a) IMDRF MDCE WG/N56FINAL:2019 Clinical Evaluation (N56:2019) Online: <https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-191010-mdce-n56.pdf>
386. International Medical Device Regulators Forum (IMDRF) (2019b) IMDRF MDCE WG/N55FINAL:2019 Clinical Evidence - Key Definitions and Concepts Online: <https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-191010-mdce-n56.pdf>
387. International Medical Device Regulators Forum (IMDRF) (2019c) IMDRF MDCE WG/N57FINAL:2019. Clinical investigation. Online: <https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-191010-mdce-n57.pdf>
388. International Medical Devices Regulators Forum (IMDRF) (2020a) IMDRF/AE WG/N43FINAL:2020 (Edition 4) IMDRF terminologies for categorized Adverse Event Reporting (AER): terms, terminology structure and codes. Online: <https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-200318-ae-terminologies-n43.pdf>

389. International Medical Devices Regulators Forum (IMDRF) (2020b) Web browser of IMDRF terminologies for Adverse Event Reporting. Online: <https://www.imdrf.org/working-groups/adverse-event-terminology>
390. International Medical Devices Regulators Forum (IMDRF) (2025a) Characterization Considerations for Medical Device Software and Software- Specific Risk. Online: <https://www.imdrf.org/documents/characterization-considerations-medical-device-software-and-software-specific-risk>
391. International Medical Devices Regulators Forum (IMDRF) (2025b) Good machine learning practice for medical device development: Guiding principles. Online: <https://www.imdrf.org/sites/default/files/2024-06/Good%20machine%20learning%20practice%20for%20medical%20device%20development%20-%20Guiding%20Principles%20DRAFT%20for%20Consultation.pdf>

## ISO

392. International standardisation organisation, ISO & International electrotechnical commission (IEC) (2008) ISO/IEC 25012:2008. Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model. Online (paywall): <https://www.iso.org/standard/35736.html>
393. International standardisation organisation, ISO & International electrotechnical commission (IEC) (2015a) ISO/IEC 25024: Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Measurement of data quality. Online (paywall): <https://www.iso.org/standard/35749.html>
394. International Standardisation Organisation, ISO (2015b) IEC 62366-1:2015. Medical devices. Part 1: Application of usability engineering to medical devices. Online (paywall): <https://www.iso.org/standard/63179.html>
395. International Standardisation Organisation, ISO (2018a) ISO/IEC 27000:2018 Information technology — Security techniques — Information security management systems — Overview and vocabulary. Online (paywall): <https://www.iso.org/standard/73906.html>
396. International Standardisation Organisation, ISO (2018b) ISO 31000:2018. Risk management — Guidelines. Online (paywall): <https://www.iso.org/standard/65694.html>
397. International Standardisation Organisation, ISO (2019) EN ISO 14971:2019, Medical devices — Application of risk management to medical devices. Online: <https://www.iso.org/standard/72704.html> (paywall).
398. International Standardisation Organisation, ISO (2020) Clinical investigation of medical devices for human subjects — Good clinical practice. Online: <https://www.iso.org/standard/71690.html> (behind paywall).
399. International Standardisation Organisation, ISO / International Electrotechnical Commission, IEC (2020) TR 29119-11:2020 (2020) Software and systems engineering — Software testing. Part 11: Guidelines on the testing of AI-based systems. Online (paywall): <https://www.iso.org/standard/79016.html>
400. International organisation for standardization, ISO (2022) Information technology — Artificial intelligence — Artificial intelligence concepts and terminology. Online: <https://www.iso.org/standard/74296.html>
401. International Standardisation Organisation, ISO (2023a) ISO/IEC 23894:2023 (2023) Information technology — Artificial intelligence — Guidance on risk management. Online (paywall): <https://www.iso.org/standard/77304.html>
402. International Standardisation Organisation, ISO (2023b) ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system. Online (paywall): <https://www.iso.org/standard/81230.html>

403. Irish data Protection Commission (2024) provides an informative website on GDPR provisions: Online: [https://www.dataprotection.ie/en/individuals/know-your-rights/right-access-information#:~:text=The%20General%20Data%20Protection%20Regulation,other%20relevant%20information%20\(as%20detailed](https://www.dataprotection.ie/en/individuals/know-your-rights/right-access-information#:~:text=The%20General%20Data%20Protection%20Regulation,other%20relevant%20information%20(as%20detailed)) (Last accessed: 2024.09.04)
404. Isensee F et al. (2021) nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. Feb;18(2):203-211. Online: doi: 10.1038/s41592-020-01008-z.
405. Islam R et al. (2024) Benchmarking Artificial Neural Network Architectures for High-Performance Spiking Neural Networks. *Sensors (Basel)*. Feb 19;24(4):1329. Online: doi: 10.3390/s24041329
406. Islam R, Weir C, Del Fiol G. (2014) Heuristics in Managing Complex Clinical Decision Tasks in Experts' Decision Making. *IEEE Int Conf Healthc Inform*. Online: doi:10.1109/ICHI.2014.32.
407. Ismail L et al. (2020) Requirements of Health Data Management Systems for Biomedical Care and Research: Scoping Review. *J Med Internet Res*. 2020 Jul 7;22(7):e17508. Online: doi: 10.2196/17508.

## J

408. Jankowski J (2014) Bioethics, Clinical. Reference Module in Biomedical Sciences, Elsevier. ISBN 9780128012383. Online: <https://doi.org/10.1016/B978-0-12-801238-3.00171-9>
409. Jenkins R & Nericcio L (2023) Evaluating Metrics for Impact Quantification. Online: <https://ssrn.com/abstract=4607297>
410. Jia H et al. (2012) Balancing human agency and object agency: an end-user interview study of the internet of things. In Proceedings of the 2012 ACM Conference on Ubiquitous Computing (pp. 1185–1188). ACM. <http://doi.org/10.1145/2370216.2370470>
411. Jiang, LY et al. (2023) Health system-scale language models are all-purpose prediction engines. *Nature* 619, 357–362 (2023). <https://doi.org/10.1038/s41586-023-06160-y>
412. Jobin A et al. (2019) The global landscape of AI ethics guidelines. *Nat Mach Intell* 1, 389–399. Online: <https://doi.org/10.1038/s42256-019-0088-2>
413. Johnson CY (2019) Racial bias in a medical algorithm favors white patients over sicker black patients. Online: <https://www.washingtonpost.com/health/2019/10/24/racial-bias-medical-algorithm-favors-white-patients-over-sicker-black-patients/>
414. Jones K (2023), AI governance and human rights: Resetting the relationship, Research Paper, London: Royal Institute of International Affairs. Online: <https://doi.org/10.55317/9781784135492>
415. Jordan J et al. (2023) Synthetic data – what, why and how? Paper commissioned by the Royal Society & Alan Turing Institute. arXiv:2205.03257. Online: <https://doi.org/10.48550/arXiv.2205.03257>

## K

416. Kaddour J. et al. (2022) Causal machine learning: a survey and open problems. arXiv:2206.15475. Online: <https://doi.org/10.48550/arXiv.2206.15475>
417. Kagerbauer SM et al. (2024) Susceptibility of AutoML mortality prediction algorithms to model drift caused by the COVID pandemic. *BMC Med Inform Decis Mak*. Feb 2;24(1):34. Online: doi: 10.1186/s12911-024-02428-z. Erratum concerning typesetting in: *BMC Med Inform Decis Mak*. 2024 Feb 19;24(1):56. doi: 10.1186/s12911-024-02454-x.
418. Kaliki S et al. (2023) Artificial intelligence and machine learning in ocular oncology: Retinoblastoma. *Indian J Ophthalmol*. Feb;71(2):424-430. doi: 10.4103/ijo.IJO\_1393\_22

419. Kaminski ME et al. (2020) Algorithmic Impact Assessments under the GDPR: Producing Multi-layered Explanations (September 18, 2019). International Data Privacy Law, 2020, forthcoming., U of Colorado Law Legal Studies Research Paper No. 19-28. Online: SSRN: <https://ssrn.com/abstract=3456224> or <http://dx.doi.org/10.2139/ssrn.3456224>
420. Kamiran F & Calders T (2009) Classifying without discriminating. 2nd International Conference on: Computer, Control and Communication. IEEE explore. Online: <https://ieeexplore.ieee.org/document/4909197?arnumber=4909197> (DOI:10.1109/IC4.2009.4909197)
421. Kamiran F & Calders T (2012) Data preprocessing techniques for classification without discrimination. *Knowl Inf Syst* 33, 1–33 Online: <https://doi.org/10.1007/s10115-011-0463-8>
422. Kamishima T et al. (2012) Fairness-Aware Classifier with Prejudice Remover Regularizer. In: Flach, P.A., De Bie, T., Cristianini, N. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2012. Lecture Notes in Computer Science(), vol 7524. Springer, Berlin, Heidelberg. Online: [https://doi.org/10.1007/978-3-642-33486-3\\_3](https://doi.org/10.1007/978-3-642-33486-3_3)
423. Kant I (1785) Grundlegung zur Metaphysik der Sitten. Felix Meiner Verlag, 1965.
424. Karches KE (2018) Against the iDoctor: why artificial intelligence should not replace physician judgment. *Theor Med Bioeth.* Apr;39(2):91–110. Online: doi: 10.1007/s11017-018-9442-3.
425. Karimi D et al. (2020) Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Med Image Anal.* Oct;65:101759. Online: doi: 10.1016/j.media.2020.101759. Epub 2020 Jun 20.
426. Kashyap P (2024) Understanding RMSProp: A Simple Guide to One of Deep Learning's Powerful Optimizers. Medium website. Online: <https://medium.com/@piyushkashyap045/understanding-rmsprop-a-simple-guide-to-one-of-deep-learning-s-powerful-optimizers-403baeed9922#:~:text=RMSProp%2C%20short%20for%20Root%20Mean,models%20dealing%20with%20sparse%20data>
427. Keil FC (2006) Explanation and understanding. *Annu Rev Psychol.* 57:227–54. doi: 10.1146/annurev.psych.57.102904.190100
428. Kelly CJ et al. (2019) Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* Oct 29;17(1):195. Online: doi: 10.1186/s12916-019-1426-2.
429. Keysers C & Gazzola V. (2014) Hebbian learning and predictive mirror neurons for actions, sensations and emotions. *Philos Trans R Soc Lond B Biol Sci.* Apr 28;369(1644):20130175. Online: doi: 10.1098/rstb.2013.0175. PMID: 24778372; PMCID: PMC4006178
430. Khalid N et al. (2023) Privacy-preserving artificial intelligence in healthcare: Techniques and applications, *Computers in Biology and Medicine* 158, 106848. Online: <https://doi.org/10.1016/j.combiomed.2023.106848>
431. Khan H et al. (2025) Chapter 21 - Transformer networks and autoencoders in genomics and genetic data interpretation: A case study, Editor(s): Khalid Raza, Deep Learning in Genetics and Genomics, Academic Press. Pages 399–423. Online: <https://doi.org/10.1016/B978-0-443-27523-4.00004-4>.
432. Khera R et al. (2023) Automation Bias and Assistive AI: Risk of Harm From AI-Driven Clinical Decision Support. *JAMA.* 2023 Dec 19;330(23):2255–2257. Online: doi: 10.1001/jama.2023.22557.
433. Kinnas M et al. (2024) Reducing Inference Energy Consumption Using Dual Complementary CNNs. arXiv:2412.01039. Online: <https://doi.org/10.48550/arXiv.2412.01039> or <https://doi.org/10.1016/j.future.2024.107606>
434. Kinsman L et al. (2010) What is a clinical pathway? Development of a definition to inform the debate. *BMC Med* 8, 31. Online: <https://doi.org/10.1186/1741-7015-8-31>
435. Kirk HR et al. (2024) The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nat Mach Intell* 6, 383–392 (2024). <https://doi.org/10.1038/s42256-024-00820-y>

436. Kiseleva A et al. (2022) Transparency of AI in Healthcare as a Multilayered System of Accountabilities: Between Legal Requirements and Technical Limitations. *Front Artif Intell.* 2022 May 30;5:879603. doi: 10.3389/frai.2022.879603
437. Klaiber M et al. (2023) The 10 most popular Concept Drift Algorithms: An overview and optimisation potentials. *Procedia Computer Science*, Volume 225, Pages 1261-1271. Online: <https://doi.org/10.1016/j.procs.2023.10.114>
438. Kleinberg et al. (2016) Inherent Trade-Offs in the Fair Determination of Risk Scores. Online: arXiv:1609.05807 or: <https://doi.org/10.48550/arXiv.1609.05807>
439. Kluge EH et al. (2018) Ethics certification of health information professionals. IMIA yearbook of medical informatics 2018. Online: <https://www.thieme-connect.de/products/ejournals/abstract/10.1055/s-0038-1641196>
440. Kluge Corrêa et al. (2023) Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance, *Patterns* 4(10): 100857. Online: <https://doi.org/10.1016/j.patter.2023.100857>
441. Kohler S (2025) Technology Federalism: U.S. States at the Vanguard of AI Governance. Carnegie endowment for international peace. Online: [https://carnegie-production-assets.s3.amazonaws.com/static/files/Kohler\\_Technology%20Federalism.pdf](https://carnegie-production-assets.s3.amazonaws.com/static/files/Kohler_Technology%20Federalism.pdf)
442. Koch C & Poggio T (1987) Artificial intelligence. In: Encyclopedia of neuroscience. Vol I. pp. 77-80. Birkhäuser, Boston.
443. Koleva-Kolarova R et al (2022) Financing and Reimbursement Models for Personalised Medicine: A Systematic Review to Identify Current Models and Future Options. *Appl Health Econ Health Policy*. Jul;20(4):501-524. Online: doi: 10.1007/s40258-021-00714-9
444. Konduri N et al. (2017) User experience analysis of an eHealth system for tuberculosis in resource-constrained settings: A nine-country comparison. *Int J Med Inform.* Jun;102:118-129. Online: doi: 10.1016/j.ijmedinf.2017.03.017
445. Kopjar, V (2021) An Overview of the Nuremberg Code, Declaration of Helsinki and Belmont Report in the Context of Promoting Ethical Global Clinical Trial Conduct.) *J Clin Res* (2021):131.
446. Kora P et al. (2022) Transfer learning techniques for medical image analysis: A review. *Biocybernetics and Biomedical Engineering*. Volume 42, Issue 1: 79-107. Online: <https://doi.org/10.1016/j.bbe.2021.11.004>
447. Korot, E et al. (2021) Predicting sex from retinal fundus photographs using automated deep learning. *Sci Rep* 11, 10286. Online: <https://doi.org/10.1038/s41598-021-89743-x>
448. Krizhevsky A et al., (2012). ImageNet classification with deep convolutional neural net- works. In: Pereira F, CJC B, Bottou L, Weinberger KQ, editors. *Advances in neural information processing systems*. 25 (NIPS 2012) Online: [https://proceedings.neurips.cc/paper\\_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf)
449. Kulesza T et al. (2013) Too much, too little, or just right? Ways explanations impact end users' mental models, in: IEEE Symposium on Visual Languages and Human-Centric Computing, VL/HCC, IEEE, Raleigh, NC, USA, 2013, pp. 3–10, <http://dx.doi.org/10.1109/vlhcc.2013.6645235>.
450. Kulesza T et al. (2015) Principles of explanatory debugging to personalize interactive machine learning, in: Proceedings of the 20th International Conference on Intelligent User Interfaces, ACM, Atlanta, Georgia, USA, pp. 126–137, <http://dx.doi.org/10.1145/2678025.2701399>.

## L

451. Lacave C & Díez FJ (2002) A review of explanation methods for Bayesian networks, *Knowl. Eng. Rev.* 17 (2) (2002) 107–127. Online: <http://dx.doi.org/10.1017/s026988890200019x>
452. Lagemann K et al. (2023) Deep learning of causal structures in high dimensions under data limitations. *Nat Mach Intell* 5, 1306–1316 (2023). Online: <https://doi.org/10.1038/s42256-023-00744-z>

453. Lamy J-B et al. (2019) Explainable artificial intelligence for breast cancer: a visual case- based reasoning approach. *Artif Intell Med.* 94:42–53. Online: doi: 10.1016/j.artmed.2019.01.001
454. Lapicque L (1907, translated and republished 2007). Recherches quantitatives sur l'excitation électrique des nerfs traitée comme une polarisation". *J. Physiol. Pathol. Gen.* 9: 620–635. Translated in English by Brunel & van Rossum, 2007.
455. Laplace de PS (1825; 1995) *Essai Philosophique sur les Probabilités*, fifth edition. Springer, NewYork. Translated by A.I. Dale (1995).
456. Laranjo L (2018) Conversational agents in healthcare: a systematic review. *J Am Med Inform Assoc.* Sep 1;25(9):1248–1258. Online: doi: 10.1093/jamia/ocy072.
457. Last JM (ed.) (2007) A dictionary of public health. Oxford University Press. Entry on “bias” online: <https://www.oxfordreference.com/display/10.1093/oi/author-it/20110803095504939#:~:text=Systematic%20distortion%20of%20results%20or,that%20influence%20conclusions%20and%20decisions> .
458. Latzer et al. (2014) The Economics of Algorithmic Selection on the Internet (October 1, 2014). In: Bauer, J. and Latzer, M. (Eds), *Handbook on the Economics of the Internet*. Cheltenham, Northampton: Edward Elgar, 395-425., Available at SSRN: <https://ssrn.com/abstract=2710399> or <http://dx.doi.org/10.2139/ssrn.2710399>
459. Laux J (2023) Institutionalised distrust and human oversight of artificial intelligence: towards a democratic design of AI governance under the European Union AI Act. *AI & SOCIETY*. Online: <https://doi.org/10.1007/s00146-023-01777-z>
460. Laux J et al. (2024) Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regul Gov.* Jan;18(1):3–32. doi: 10.1111/rego.12512. Epub 2023 Feb 6. Online: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10903109/pdf/REGO-18-3.pdf>
461. Lavazza A (2016) Free Will and Neuroscience: From Explaining Freedom Away to New Ways of Operationalizing and Measuring It. *Front Hum Neurosci.* Jun 1;10:262. Online: doi: 10.3389/fnhum.2016.00262
462. Lazzaro D et al. (2022) Minimizing Energy Consumption of Deep Learning Models by Energy-Aware Training. *arXiv:2307.00368*. Online: <https://doi.org/10.48550/arXiv.2307.00368>
463. Lee A & Leung S (2014) Health Outcomes. In: Michalos, A.C. (eds) *Encyclopedia of Quality of Life and Well-Being Research*. Springer, Dordrecht. [https://doi.org/10.1007/978-94-007-0753-5\\_1251](https://doi.org/10.1007/978-94-007-0753-5_1251)
464. Lee CS & Lee AY (2020) Clinical applications of continual learning machine learning. *Lancet Digit Health.* Jun;2(6):e279–e281. Online: doi: 10.1016/S2589-7500(20)30102-3
465. Leena AI labs (2024) What is continuous learning? Online: <https://www.leena.ai/ai-glossary/continuous-learning>
466. LeLagadec D et al. (2024), Navigating the impact of artificial intelligence on our healthcare workforce. *J Clin Nurs.* 33: 2369–2370. Online: <https://doi.org/10.1111/jcn.17191>
467. Lenatti M et al. (2023) Characterization of Synthetic Health Data Using Rule-Based Artificial Intelligence Models. *IEEE J Biomed Health Inform.* Aug;27(8):3760–3769. Online: doi: 10.1109/JBHI.2023.3236722
468. Leslie D – The Alan Turing Institute (2019) Understanding artificial intelligence ethics and safety. A guide for the responsible design and implementation of AI systems in the public sector. Online: [https://www.turing.ac.uk/sites/default/files/2019-06/understanding\\_artificial\\_intelligence\\_ethics\\_and\\_safety.pdf](https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf)
469. Lefkowitz M (2019) Professor's perceptron paved the way for AI – 60 years too soon. Cornell Chronicle. Online: <https://news.cornell.edu/stories/2019/09/professors-perceptron-paved-way-ai-60-years-too-soon>

470. Li Z (2022) Efficiency analysis of artificial vs. Spiking Neural Networks on FPGAs, Journal of Systems Architecture, Volume 133, 102765. Online: <https://doi.org/10.1016/j.sys-arc.2022.102765>.
471. Li VJ et al. (2023) A Guide for the Multiplexed: The Development of Visual Feature Maps in the Brain, Neuroscience, Volume 508: 62–75. Online: <https://doi.org/10.1016/j.neuroscience.2022.07.026>
472. Libet B et al. (1983) Time of conscious intention to act in relation to onset of cerebral activities (readiness-potential): the unconscious initiation of a freely voluntary act. Brain 106, 623–642. doi: 10.1093/brain/106.3.623
473. Linardatos P et al. (2021) Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy 23(1), 18. Online: <https://doi.org/10.3390/e23010018>
474. Lindblad KE & Bloch Veiberg C (2020) Human rights impact assessment of digital activities. The Danish Institute for Human Rights. <https://www.humanrights.dk/publications/human-rights-impact-assessment-digital-activities>
475. Lipton ZC (2017) The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue, vol. 16, no. 3, pp. 31–57, 2018. arXiv:1606.03490. Online: <https://doi.org/10.48550/arXiv.1606.03490> (Last accessed: 2024.08.16)
476. Little C et al., (2021) Generative Adversarial Networks for Synthetic Data Generation: A Comparative Study. arXiv. Online: <https://doi.org/10.48550/arXiv.2112.01925>
477. Liu et al. (2020) Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med 26, 1364–1374 (2020). Online: <https://doi.org/10.1038/s41591-020-1034-x>
478. Liu H et al. (2022) Artificial Intelligence model development and validation. In: Matheny et al. (eds) (2022) Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril. Washington, DC: National Academy of Medicine: The National Academies Press. Online: <https://doi.org/10.17226/27111>
479. Liu W et al. (2024) Aligning Large Language Models with Human Preferences through Representation Engineering. arXiv. Online: arXiv:2312.15997
480. Løgstrup KE (1956) The ethical demand. University of Notre Dame Press, Indiana, USA, 1997. Online: <https://undpress.nd.edu/9780268009342/the-ethical-demand/>
481. Lombrozo T (2012) Explanation and abductive inference. In: Oxford handbook of thinking and reasoning, p. 260–276. Online: <https://academic.oup.com/edited-volume/34559>
482. London AJ (2019) Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. The Hastings Center Report 49(1):15–21. Online: <https://doi.org/10.1002/hast.973>
483. Lou Y, Caruana R, Gehrke J (2012) Intelligible models for classification and regression. KDD, 2012. Online: <https://dl.acm.org/doi/10.1145/2339530.2339556>
484. Lundgren B & Möller N (2019) Defining Information Security. Sci Eng Ethics 25, 419–441. Online: <https://doi.org/10.1007/s11948-017-9992-1>
485. Lupiáñez-Villanueva et al. (2022) Study on health data, digital health and artificial intelligence in healthcare. Study commissioned by EU Commission. Online: [https://health.ec.europa.eu/publications/study-health-data-digital-health-and-artificial-intelligence-healthcare\\_en](https://health.ec.europa.eu/publications/study-health-data-digital-health-and-artificial-intelligence-healthcare_en) (Last accessed: 2024.09.18)

## M

486. MachineLearningMastery (website) (2024) 5 Challenges in Machine Learning Adoption and How to Overcome Them. Online: <https://machinelearningmastery.com/5-challenges-in-machine-learning-adoption-and-how-to-overcome-them/> (Last accessed: 2024.11.14)

487. Madary M & Metzinger TK (2016) Real Virtuality: A Code of Ethical Conduct. Recommendations for Good Scientific Practice and the Consumers of VR-Technology. *Front. Robot. AI*, 19, Sec. Virtual Environments, Volume 3. Online: <https://doi.org/10.3389/frobt.2016.0000>
488. Manyara AM et al. (2024) Reporting of surrogate endpoints in randomised controlled trial reports (CONSORT-Surrogate): extension checklist with explanation and elaboration. *BMJ*. 2024 Jul 9;386:e078524. Online: doi: 10.1136/bmj-2023-078524
489. Mardziel P (2021) Drift in machine learning. Online: <https://towardsdatascience.com/drift-in-machine-learning-e49df46803a>
490. Marewski JN & Gigerenzer G (2012) Heuristic decision making in medicine. *Dialogues Clin Neurosci*. 2012 Mar;14(1):77-89. Online: doi: 10.31887/DCNS.2012.14.1/jmarewski.
491. Martens D et al., (2007) Comprehensible credit scoring models using rule extraction from support vector machines, *European J. Oper. Res.* 183 (3) 1466–1476. Online: <http://dx.doi.org/10.1016/j.ejor.2006.04.051>
492. Martinez-Martin N (2018) Is It Ethical to Use Prognostic Estimates from Machine Learning to Treat Psychosis? *AMA J Ethics*. 20(9):E804-811. Online: doi: 10.1001/amaethics.2018.804.
493. Matheny et al. (eds) (2022) Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril. Washington, DC: National Academy of Medicine: The National Academies Press. Online: <https://doi.org/10.17226/27111>
494. Mattioli M, Cinà AE, Pelillo M(2024) Understanding XAI Through the Philosopher's Lens: A Historical Perspective. arXiv:2407.18782v1
495. Mayer D (2009) Essential evidence-based medicine. Cambridge University Press.
496. Mayer RC et al. (1995). An integrative model of organisational trust. *Academy of Management Review*, 20(3), 709–734. Online: <https://www.jstor.org/stable/258792?origin=crossref&seq=4> (Last accessed: 2024.09.16)
497. McCarthy J, Minsky ML, Rochester N, Shannon CE (1955; reprinted in 2006). A proposal for the Dartmouth summer research project on artificial intelligence, August 21, 1955. *AI Magazine / Archives*, Vol 27(4) 2006. Online: <https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1904>
498. McCoy LG et al. (2022) Believing in black boxes: machine learning for healthcare does not need explainability to be evidence-based. *J Clin Epidemiol*. Feb;142:252-257. Online: doi: 10.1016/j.jclinepi.2021.11.001. Epub 2021 Nov 5. PMID: 34748907.
499. McCulloch WS & Pitts W (1943, reprinted 1990): A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biology* Vol 52(1/2):99-115. Online: <https://www.cs.cmu.edu/~epxing/Class/10715/reading/McCulloch.and.Pitts.pdf>
500. McGilvray D (2008) Executing Data Quality Projects Ten Steps to Quality Data and Trusted Information. Elsevier, Amsterdam.
501. McGregor L et al. (2019), International Human Rights Law as a Framework for Algorithmic Accountability. *International & Comparative Law Quarterly*, 68(2), April 2019, pp. 309–43. Online: <https://doi.org/10.1017/S0020589319000046>
502. McNair D, Price WN (2019) Health care AI: Law, regulation, and policy. In: Matheny M, Thadaney Israni S, Ahmed M, Whicher D, editors. Artificial intelligence in health care: The hope, the hype, the promise, the peril. Washington DC: National Academy of Medicine; 2019. Online: <https://nap.nationalacademies.org/read/27111/chapter/1>
503. Meng XL (2020) Reproducibility, replicability, and reliability. *Harvard data science review*. Issue 2.4. Online: <https://hdsr.mitpress.mit.edu/pub/hn51kn68/release/4> (Last accessed: 2024.08.22).
504. Mennella C et al. (2024) Ethical and regulatory challenges of AI technologies in healthcare: A narrative review. *Heliyon*. 2024 Feb 15;10(4):e26297. Online: doi: 10.1016/j.heliyon.2024.e26297.
505. Microsoft (2025) Model fine-tuning concepts. Online: <https://learn.microsoft.com/en-us/windows/ai/fine-tuning>

506. Microsoft Ignite (2025) A list of metrics for evaluating LLM-generated content. Online: <https://learn.microsoft.com/en-us/ai/playbook/technology-guidance/generative-ai/working-with-langs/evaluation/list-of-eval-metrics>
507. Miller JB & Gelman A (2018) Laplace's Theories of Cognitive Illusions, Heuristics, and Biases (December 1, 2018). Available at SSRN: <https://ssrn.com/abstract=3149224> or <http://dx.doi.org/10.2139/ssrn.3149224> and published in Statistical Science 2020, Vol. 35, No. 2, 159–170. Online: <https://doi.org/10.1214/19-STS696>
508. Miller T (2019) Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence 267:1–38. Online: <https://doi.org/10.1016/j.artint.2018.07.007>
509. Miller A (2022) Open Data Science.com. Guidelines for Choosing an Optimizer and Loss Functions When Training Neural Networks. Online: <https://opendatascience.com/guidelines-for-choosing-an-optimizer-and-loss-functions-when-training-neural-networks/>
510. Ming, J. (2022) Health technology assessment of medical devices: current landscape, challenges, and a way forward. Cost Eff Resour Alloc 20, 54. Online: <https://doi.org/10.1186/s12962-022-00389-6>
511. Minsky ML (1969) Semantic information processing. Cambridge, MA: MIT Press. Cited from: JRC technical report: AI Watch. Defining artificial intelligence. EU 2020. Online: doi: 10.2760/38273
512. Minsky ML & Papert S (1969) Perceptrons: an introduction to computational geometry. MIT press, Cambridge, Mass.
513. Mishra M (2023) The Learning Rate: A Hyperparameter That Matters. Medium website. Online: <https://mohitmishra786687.medium.com/the-learning-rate-a-hyperparameter-that-matters-b2f3b68324ab>
514. Mishra S et al. (2024) Reliability, Resilience and Human Factors Engineering for Trustworthy AI Systems. Online: <https://doi.org/10.48550/arXiv.2411.08981>
515. MIT technology review (2024) Artificial intelligence is creating a new colonial world order. Online: <https://www.technologyreview.com/2022/04/19/1049592/artificial-intelligence-colonialism/>
516. Mittelstadt BD et al. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2). <https://doi.org/10.1177/2053951716679679>
517. Mittelstadt BD (2019) Principles alone cannot guarantee ethical AI. Nat Mach Intell 1, 501–507. Online: <https://doi.org/10.1038/s42256-019-0114-4>
518. Mittelstadt BD (2021) The impact of artificial intelligence on the doctor-patient relationship. Report commissioned by the Council of Europe's Steering Committee for Human Rights in the fields of biomedicine and health (CDBIO). Online: <https://rm.coe.int/inf-2022-5-report-impact-of-ai-on-doctor-patient-relations-e/1680a68859>
519. Mittermaier M et al. (2023) Bias in AI-based models for medical applications: challenges and mitigation strategies. npj Digit. Med. 6, 113. Online: <https://doi.org/10.1038/s41746-023-00858-z>
520. Mollaeeefar M & Ranise S (2023) Identifying and quantifying trade-offs in multi-stakeholder risk evaluation with applications to the data protection impact assessment of the GDPR, Computers & Security, Volume 129, <https://doi.org/10.1016/j.cose.2023.103206>
521. Mongan J, et al. (2020) Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell 2020;2: e200029. <https://doi.org/10.1148/rhai.2020200029>
522. Moolla Y (2024) Postcodes: hidden proxies for protected attributes. Online: <https://www.linkedin.com/pulse/postcodes-hidden-proxies-protected-attributes-yusuf-moolla-gf0oc/>
523. Moons Karel GM, et al. "Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration." Annals of internal medicine

- 162.1 (2015): W1-W73. Online: [https://www.acpjournals.org/doi/full/10.7326/M14-0698?rfr\\_dat=cr\\_pub+Opubmed&url\\_ver=Z39.88-2003&rfr\\_id=ori%3Arid%3Acrossref.org](https://www.acpjournals.org/doi/full/10.7326/M14-0698?rfr_dat=cr_pub+Opubmed&url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org)
524. Moor M et al. (2023) Foundation models for generalist medical artificial intelligence. *Nature*. 2023 Apr;616(7956):259-265. doi: 10.1038/s41586-023-05881-4
525. Morley J (2020a). From what to how. An overview of AI ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics*, 26, 2141–2168. Online: <https://link.springer.com/article/10.1007/s11948-019-00165-5>
526. Morley J et al. (2020b) The ethics of AI in health care: A mapping review. *Soc. Sci Med.* 260:113172. doi: 10.1016/j.socscimed.2020.113172.
527. Mosier KL & JL Skitka (1996) Human Decision Makers and Automated Decision Aids: Made for Each Other? Chapter 10, pp. 201-220. In: "Automation and Human Performance: Theory and Applications". Online (via Research Gate): [https://www.researchgate.net/publication/230601064\\_Human\\_Decision\\_Makers\\_and\\_Automated\\_Decision\\_Aids\\_Made\\_for\\_Each\\_Other](https://www.researchgate.net/publication/230601064_Human_Decision_Makers_and_Automated_Decision_Aids_Made_for_Each_Other)
528. Mosquera L et al. (2023) A method for generating synthetic longitudinal health data. *BMC Med Res Methodol*. Mar 23;23(1):67. Online: doi: 10.1186/s12874-023-01869-w
529. Muldoon J, Wu BA (2023) Artificial Intelligence in the Colonial Matrix of Power. *Philos. Technol.* 36, 80 (2023). <https://doi.org/10.1007/s13347-023-00687-8>
530. Müller CM & Griesinger CB (1998) Tissue plasminogen activator mediates reverse occlusion plasticity in visual cortex. *Nat Neurosci*. May;1(1):47-53. Online: doi: 10.1038/248. PMID: 10195108
531. Müller D et al. (2022) Towards a guideline for evaluation metrics in medical image segmentation. *BMC Res Notes*. 2022 Jun 20;15(1):210. doi: 10.1186/s13104-022-06096-y.
532. Müller H et al. (2022) Explainability and causability for artificial intelligence-supported medical image analysis in the context of the European In Vitro Diagnostic Regulation. *N Biotechnol*. Sep 25;70:67-72. Online: doi: 10.1016/j.nbt.2022.05.002.
533. Murdoch WJ et al. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, 116(44), 22071- 22080. Online: <https://www.pnas.org/doi/full/10.1073/pnas.1900654116>
534. Mutasa S et al. (2021) Understanding artificial intelligence based radiology studies: What is overfitting? *Clin Imaging*. 2020 Sep;65:96-99. doi: 10.1016/j.clinimag.2020.04.025.
535. Myllyaho et al. (2021) Systematic literature review of validation methods for AI systems. *Journal of Systems and Software*, Volume 181, 2021. Online: <https://doi.org/10.1016/j.jss.2021.111050>.

## N

536. Nadeem M et al. (2020). Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv*. Online: arXiv:2004.09456.
537. Naeem MA (2024) How an algorithm transitions into a model in machine learning. Online: <https://www.linkedin.com/pulse/how-algorithm-transitions-model-machine-learning-muhammad-adil-naeem-erwcc/> (Last accessed: 2025.02.17)
538. Najafian S (2022) A theory of cortical map formation in the visual brain. *Nat Commun*. 2022 Apr 28;13(1):2303. Online: doi: 10.1038/s41467-022-29433-y
539. Najafzadeh M et al. (2015) A unified framework for classification of methods for benefit-risk assessment. *Value Health*. 2015 Mar;18(2):250-9. doi: 10.1016/j.jval.2014.11.001
540. Naqvi H & L'Esperance V (2024) Tackling bias in medical devices: the Equity in Medical Devices Independent Review is welcome, but could have gone further. *BMJ*. Mar 12;384:q620. Online: doi: 10.1136/bmj.q620

541. Narmadha K & Varalakshmi P (2022) Federated Learning in Healthcare: A Privacy Preserving Approach. Stud Health Technol Inform. May 25;294:194-198. doi: 10.3233/SHTI220436. PMID: 35612055.
542. National Academy of Sciences (2009) Reproducibility and replicability in science. Online: [https://www.ncbi.nlm.nih.gov/books/NBK547537/pdf/Bookshelf\\_NBK547537.pdf](https://www.ncbi.nlm.nih.gov/books/NBK547537/pdf/Bookshelf_NBK547537.pdf)

## NIST

543. National Institute of Standards and Technology (NIST) (2004) Computer Security Incident Handling Guide. Special Publication (SP) 800-61. <https://doi.org/10.6028/NIST.SP.800-61>
544. National Institute of Standards and Technology (NIST) (2014) Guidelines for media sanitization. Online: <https://csrc.nist.gov/pubs/sp/800/88/r1/final>
545. National Institute of Standards and Technology (NIST) (2017a) An introduction to information security. Special Publication 800-12 Revision 1. Online: <https://nvlpubs.nist.gov/nistpubs/Special-Publications/NIST.SP.800-12r1.pdf>
546. National Institute of Standards and Technology (NIST) (2017b) An Introduction to Privacy Engineering and Risk Management in Federal Systems. Internal Report 8062. Online: <https://doi.org/10.6028/NIST.IR.8062>
547. National Institute of Standards and Technology (NIST) (2021a). NISTIR 8312. Four principles of explainable artificial intelligence. Online: <https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8312.pdf> (Last accessed: 2024.08.16).
548. National Institute of Standards and Technology (NIST) (2021b). NISTIR 8367. Psychological foundations of explainability and interpretability in AI. Online: <https://www.nist.gov/publications/psychological-foundations-explainability-and-interpretability-artificial-intelligence> (Last accessed: 2024.08.16)
549. National Institute for Standards and Technology (NIST) (2022) Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. Online: [https://tsapps.nist.gov/publication/get\\_pdf.cfm?pub\\_id=934464](https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=934464) (Last accessed: 2024.09.02)
550. National Institute of Standards and Technology (NIST) (2023) NIST AI 100-1. Artificial intelligence risk management framework. Online: <https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf> (Last accessed: 2024-09-02).
551. National Institute of Standards and Technology (NIST) (2024a) Special Publication 1800-28B. Online: <https://www.nccoe.nist.gov/sites/default/files/2024-02/dc-ip-nist-sp-1800-28b-final.pdf>
552. National Institute for Standards and Technology (NIST) (2024b) Glossary of the NIST's computer security resource center. Online: [https://csrc.nist.gov/glossary/term/system\\_administrator](https://csrc.nist.gov/glossary/term/system_administrator)
553. National Information Security Standardization Technical Committee (TC260) of the People's Republic of China (2024) AI safety governance framework. Online: <https://www.tc260.org.cn/upload/2024-09-09/1725849192841090989.pdf>
554. National library of medicine (2024) Algorithmic bias. Online: <https://www.ncbi.nlm.nih.gov/guides/data-thesaurus/algorithmic-bias>
555. Nazer LH et al. (2023) Bias in artificial intelligence algorithms and recommendations for mitigation. PLOS Digit Health. Jun 22;2(6):e0000278. Online: doi: 10.1371/journal.pdig.0000278
556. Nerella S et al. (2024) Transformers and large language models in healthcare: A review. Artif Intell Med. Aug;154:102900. Online: doi: 10.1016/j.artmed.2024.102900.
557. Nerincx N & Lindenberg J (2005) Integrating Human Factors and Artificial Intelligence in the Development of Human-Machine Cooperation. Proceedings of the 2005 International Conference on Artificial Intelligence, ICAI 2005, Las Vegas, Nevada, USA, June 27-30, 2005, Volume 1. Online: [https://www.researchgate.net/publication/220834351\\_Integrating\\_Human\\_Factors\\_and\\_Artificial\\_Intelligence\\_in\\_the\\_Development\\_of\\_Human-Machine\\_Cooperation](https://www.researchgate.net/publication/220834351_Integrating_Human_Factors_and_Artificial_Intelligence_in_the_Development_of_Human-Machine_Cooperation)

558. Nightfall AI (2024) Model integrity verification. Online: <https://www.nightfall.ai/ai-security-101/model-integrity-verification#:~:text=AI%20its%20core%2C%20Model%20Integrity,where%20its%20outputs%20are%20trustworthy>.
559. Ng MY et al. (2022) The AI life cycle: a holistic approach to creating ethical AI for health decisions. *Nat Med.* 2022 Nov;28(11):2247-2249. doi: 10.1038/s41591-022-01993-y
560. Niehaus et al. (2022) An occupational safety and health perspective on human in control and AI. *Frontiers in Artificial Intelligence*, Vol. 5. Online: <https://doi.org/10.3389/frai.2022.868382>
561. Nithya Sambasivan & Rajesh Veeraraghavan (2022) The Deskilling of Domain Expertise in AI Development. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 587, 1–14. Online: <https://doi.org/10.1145/3491102.3517578>
562. NITI Aayog (2018) National strategy for artificial intelligence. <https://www.niti.gov.in/sites/default/files/2023-03/National-Strategy-for-Artificial-Intelligence.pdf>
563. Noller J (2020) Heautonomy. Schiller on freedom of the will. *Eur J Philos.* 2021; 29: 339–353. Online: <https://doi.org/10.1111/ejop.12576>
564. NordLayer (2024) How to prevent unauthorized access: 10 best practices. Online: <https://nord-layer.com/blog/how-to-prevent-unauthorized-access/> (Last accessed: 2024.09.04)
565. Norgeot B. et al. (2020) Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med* 26, 1320–1324 (2020). <https://doi.org/10.1038/s41591-020-1041-y>

## 0

566. Oasis Open (2024) OASIS Collaborative Automated Course of Action Operations (CACAO) for Cyber Security. Online: <https://groups.oasis-open.org/communities/tc-community-home2?CommunityKey=b75cccb8-adc6-4de5-8b99-018dc7d322b6>
567. Obermeyer Z et al. (2019) Dissecting racial bias in an algorithm used to manage the health of populations. *Science.* Oct 25;366(6464):447-453. Online: doi: 10.1126/science.aax2342.
568. Okolo C (2024) The Brookings Institution: AI in the Global South: Opportunities and challenges towards more inclusive governance. Online: <https://www.brookings.edu/articles/ai-in-the-global-south-opportunities-and-challenges-towards-more-inclusive-governance/>
569. Onetrust (2024) China: TC260 releases AI safety governance framework. Online: <https://www.onetrust.com/blog/chinas-tc260-releases-ai-safety-governance-framework/>
570. Open encyclopedia of cognitive science. Entry: causal learning. Online: <https://oecs.mit.edu/pub/i1om74mo/release/1> (Last accessed: 2024.08.14).

## OECD

571. Organisation for economic cooperation and development (OECD) (2016) Recommendation of the Council on Health Data Governance, OECD/LEGAL/0433. Online: <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0433>
572. Organisation for economic cooperation and development (OECD) (2019a; amended 2024) Recommendation of the Council on Artificial Intelligence. Online: <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>
573. Organisation for economic cooperation and development (OECD) with the German Federal Ministry of Labour and Social affairs (2019b) OECD framework for the classification of AI systems. Series: OECD digital economy papers. Online: [https://www.oecd.org/en/publications/2022/02/oecd-framework-for-the-classification-of-ai-systems\\_336a8b57.html](https://www.oecd.org/en/publications/2022/02/oecd-framework-for-the-classification-of-ai-systems_336a8b57.html) (Last accessed: 2024.08.13)

574. Organisation for economic cooperation and development (OECD) (2021a) OECD.AI, powered by EC/OECD (2021), database of national AI policies. Online: <https://oecd.ai> (Last accessed: 2025.01.25)
575. Organisation for economic cooperation and development (OECD) (2021b) Tools for trustworthy AI. A framework to compare implementation tools for trustworthy AI systems. OECD Digital Economy Papers, No. 312, OECD Publishing, Paris Online: <https://doi.org/10.1787/008232ec-en>.
576. Organisation for economic cooperation and development (OECD) (2022) OECD Framework for the Classification of AI systems. Online: [https://www.oecd.org/en/publications/oecd-framework-for-the-classification-of-ai-systems\\_cb6d9eca-en.html](https://www.oecd.org/en/publications/oecd-framework-for-the-classification-of-ai-systems_cb6d9eca-en.html)
577. Organisation for economic cooperation and development (OECD) (2024a) Council – Report on the implementation of the OECD recommendation on artificial intelligence. Document code: C/MIN(2024)17. Online: [https://one.oecd.org/document/C/MIN\(2024\)17/en/pdf](https://one.oecd.org/document/C/MIN(2024)17/en/pdf)
578. Organisation for economic cooperation and development (OECD) (2024b) AI system definition update. Online: <https://oecd.ai/en/wonk/ai-system-definition-update>
579. Organisation for economic cooperation and development (OECD) (2024c) AI in health. Huge potential, huge risks. Online: [https://www.oecd.org/en/publications/ai-in-health\\_2f709270-en.html](https://www.oecd.org/en/publications/ai-in-health_2f709270-en.html) (Last accessed: 2024.08.25)
580. Organisation for economic cooperation and development (OECD) (2024d) Collective action for responsible AI in health. OECD ARTIFICIAL INTELLIGENCE PAPERS. January 2024 No. 10. Online: [https://www.oecd.org/en/publications/collective-action-for-responsible-ai-in-health\\_f2050177-en.html](https://www.oecd.org/en/publications/collective-action-for-responsible-ai-in-health_f2050177-en.html)
581. Organisation for economic cooperation and development (OECD) (2024e) AI, data governance and privacy. Online: [https://www.oecd.org/en/publications/ai-data-governance-and-privacy\\_2476b1a4-en.html](https://www.oecd.org/en/publications/ai-data-governance-and-privacy_2476b1a4-en.html)
582. O'Rourke B et al. (2020) The new definition of health technology assessment: A milestone in international collaboration. International Journal of Technology Assessment in Health Care. 36(3):187-190. doi:10.1017/S0266462320000215
583. Orphanet (2024) About orphan drugs. Online: <https://www.orpha.net/en/other-information/about-orphan-drugs?stapage=what> (Last accessed: 2024.09.20).
584. Osaro M (2023) Confidentiality, Integrity, and Availability in Network Systems: A Review of Related Literature. International Journal of Innovative Science and Research Technology (IJISRT), www.ijisrt.com. ISSN - 2456-2165 , PP :- 1946-1955.DOI: <https://doi.org/10.5281/zenodo.10464076>

## P

585. Page M J et al. (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews BMJ 2021; 372 :n71 doi:10.1136/bmj.n71
586. Paleyes A (2022) Challenges in Deploying Machine Learning: A Survey of Case Studies. ACM Comput. Surv. 55, 6, Article 114 (June 2023), 29 pages. <https://doi.org/10.1145/3533378>
587. Panch T et al. (2019) Artificial intelligence and algorithmic bias: implications for health systems. J Glob Health. Dec;9(2):010318. Online: doi: 10.7189/jogh.09.020318.
588. Panigutti C et al. (2023) The role of explainable AI in the context of the AI Act. FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, pages 1139 – 1150. Online: <https://doi.org/10.1145/3593013.3594069>
589. Paparova D et al. (2023) Data governance spaces: The case of a national digital service for personal health data, Information and Organisation, Volume 33, Issue 1. Online: <https://doi.org/10.1016/j.infoandorg.2023.100451>
590. Papernot N et al. (2018) Scalable private learning with PATE. Online: arXiv:1802.08908

591. Parasuraman R & Manzey DH (2010). Complacency and bias in human use of automation: an attentional integration. *Hum Factors*. Jun;52(3):381-410. Online: doi: 10.1177/0018720810376055. PMID: 21077562.
592. Parikh RB (2022) Performance Drift in a Mortality Prediction Algorithm during the SARS-CoV-2 Pandemic. medRxiv [Preprint]. Mar 1:2022.02.28.22270996. doi: 10.1101/2022.02.28.22270996. Update in: *J Am Med Inform Assoc*. 2023 Jan 18;30(2):348-354. doi: 10.1093/jamia/ocac221. PMID: 35262088; PMCID: PMC8902871.
593. Park SH & Han K (2018). Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology* 286:800809. Online: DOI: 10.1148/radiol.2017171920
594. Park SH et al. (2021) Key Principles of Clinical Validation, Device Approval, and Insurance Coverage Decisions of Artificial Intelligence. <https://doi.org/10.3348/kjr.2021.0048>
595. Park SH (2023) How to Determine If One Diagnostic Method, Such as an Artificial Intelligence Model, is Superior to Another: Beyond Performance Metrics. *Korean J Radiol*. 2023 Jul;24(7):601-605. Online: doi: 10.3348/kjr.2023.0448
596. Parthasarathy VB et al., (2024) The Ultimate Guide to Fine-Tuning LLMs from Basics to Breakthroughs: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities. arXiv:2408.13296. Online: <https://doi.org/10.48550/arXiv.2408.13296>
597. Peabody et al. (2014) New thinking on clinical utility: hard lessons for molecular diagnostics. *Am J Manag Care*. Online: <https://pubmed.ncbi.nlm.nih.gov/25365750/>
598. Petsiuk V, Das A & Saenko K (2018). RISE: Randomized Input Sampling for Explanation of Black-box Models. In BMVC, June 2018. Available also on arXiv. Online: <https://doi.org/10.48550/arXiv.1806.07421>
599. Pew Research Centre (2023) 60% of Americans Would Be Uncomfortable With Provider Relying on AI in Their Own Health Care. Online: <https://www.pewresearch.org/science/2023/02/22/60-of-americans-would-be-uncomfortable-with-provider-relying-on-ai-in-their-own-health-care/> (Last accessed: 2024.09.18)
600. Pezel T et al. (2023) AI-Based Fully Automated Left Atrioventricular Coupling Index as a Prognostic Marker in Patients Undergoing Stress CMR. *JACC Cardiovasc Imaging*. 2023 Oct;16(10):1288-1302. Online: doi: 10.1016/j.jcmg.2023.02.015.
601. PHCAS (2025) PANACEA healthcare cybersecurity advisory services. Online: <https://panaceare-search.eu/about-panacea-and-phcas>
602. Pinar Saygin A et al (2000) Turing Test: 50 Years Later. *Minds and Machines* 10, 463-518. Online: <https://doi.org/10.1023/A:1011288000451>
603. Pinaya, W. et al (2016) Using deep belief network modelling to characterize differences in brain morphometry in schizophrenia. *Sci Rep* 6, 38897. Online: <https://doi.org/10.1038/srep38897>
604. Plass M et al. (2023) Explainability and causability in digital pathology. *J Pathol Clin Res*. 2023 Jul;9(4):251-260. Online: doi: 10.1002/cjp2.322.
605. Porter, ME (1998) *The Competitive Advantage: Creating and Sustaining Superior Performance*. NY: Free Press, 1985. (Republished with a new introduction, 1998.)
606. Pöyhönen M (2000) POLSSS: surveying stakeholders about acceptability of risks and system changes, Safety Science, Volume 35, Issues 1–3, pages 123-137. Online: [https://doi.org/10.1016/S0925-7535\(00\)00027-8](https://doi.org/10.1016/S0925-7535(00)00027-8)
607. 'Prognosisresearch' website (2025) Improving prognosis and prediction research in healthcare. The website provides various tools, including the PROGRESS framework. Online: <https://www.prognosisresearch.com/>
608. Pruijt H (2006) Social Interaction With Computers: An Interpretation of Weizenbaum's ELIZA and Her Heritage. *Social science computer review* 24, No. 4: 516-523; Online: <https://www.re>

- [https://searchgate.net/publication/240711116\\_Social\\_Interaction\\_With\\_Computers\\_An\\_Interpretation\\_of\\_Weizenbaum's\\_ELIZA\\_and\\_Her\\_Heritage/](https://searchgate.net/publication/240711116_Social_Interaction_With_Computers_An_Interpretation_of_Weizenbaum's_ELIZA_and_Her_Heritage/)
609. Purdue University (2024) Glossary: Data colonialism. Online: <https://purdue.edu/critical-data-studies/collaborative-glossary/data-colonialism.php#~:text=The%20definition%20of%20E2%80%9Cdata%20colonialism,by%20their%20users%20and%20citizens>
610. Putnam H (1988) Representation and reality. Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. Online (paywall): <https://mitpress.mit.edu/9780262660747/representation-and-reality/>
611. PWC - PriceWaterhouseCoopers (2019) A practical guide to Responsible Artificial Intelligence (AI). Online: <https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai/responsible-ai-practical-guide.pdf> (Last accessed: 2024.07.01)

## Q

612. Quinonero-Candela J et al. (eds.) (2008) Dataset shift in machine learning. Neural information processing series. MIT press. <https://mitpress.mit.edu/9780262545877/dataset-shift-in-machine-learning/>

## R

613. Radclyffe C (2023) The assessment list for trustworthy artificial intelligence: A review and recommendations. *Front Artif Intell.* 2023 Mar 9;6:1020592. Online: doi: 10.3389/frai.2023.1020592.
614. Rademakers FE et al. (2025) CORE-MD clinical risk score for regulatory evaluation of artificial intelligence-based medical device software. *NPJ Digit Med.* 2025 Feb 6;8(1):90. doi: 10.1038/s41746-025-01459-8
615. Rahmani K et al. (2023) Assessing the effects of data drift on the performance of machine learning models used in clinical sepsis prediction. *Int J Med Inform.* 173:104930. doi: 10.1016/j.ijmedinf.2022.104930.
616. Rajkomar A et al. (2018) Ensuring Fairness in Machine Learning to Advance Health Equity. *Ann Intern Med.* Dec 18;169(12):866-872. Online: doi: 10.7326/M18-1990. Epub 2018 Dec 4.
617. Rajpurkar P et al. (2022) AI in health and medicine. *Nat Med* 28, 31–38. Online: <https://doi.org/10.1038/s41591-021-01614-0>
618. Rajaraman S et al. (2021) Novel loss functions for ensemble-based medical image classification. *PLoS One.* Dec 30;16(12):e0261307. Online: doi: 10.1371/journal.pone.0261307.
619. Rami A et al. (2023) Digital-care in next generation networks: Requirements and future directions. *Computer Networks* 224, 109599. Online: <https://doi.org/10.1016/j.comnet.2023.109599>.
620. Ranard BL et al. (2024) Minimizing bias when using artificial intelligence in critical care medicine, *Journal of Critical Care*, Volume 82, 154796, Online: <https://doi.org/10.1016/j.jcrc.2024.154796>.
621. Rasmy L et al. (2021) Simple Recurrent Neural Networks is all we need for clinical events predictions using EHR data. *arXiv.* Online: <https://doi.org/10.48550/arXiv.2110.00998>
622. Rawte V et al. (2023) The Troubling Emergence of Hallucination in Large Language Models - An Extensive Definition, Quantification, and Prescriptive Remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2541–2573, Singapore. Association for Computational Linguistics. Online: <https://aclanthology.org/2023.emnlp-main.155/>
623. Reddy GT et al. (2020a) Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis. *Evol. Intel.* 13, 185–196. Oline: <https://doi.org/10.1007/s12065-019-00327-1>

624. Reddy S et al. (2020b) A governance model for the application of AI in health care. *J Am Med Inform Assoc.* Mar 1;27(3):491-497. Online: doi: 10.1093/jamia/ocz192.
625. Reddy S (2024) Generative AI in healthcare: an implementation science informed translational path on application, integration and governance. *Implementation Sci* 19, 27. Online: <https://doi.org/10.1186/s13012-024-01357-9>
626. Reguero SM et al. (2025) Energy-efficient neural network training through runtime layer freezing, model quantization, and early stopping, *Computer Standards & Interfaces*, Volume 92, 103906. Online: <https://doi.org/10.1016/j.csi.2024.103906>
627. Regulation (EU) 2017/745 on medical devices (MDR) Online: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32017R0745>
628. Reina V & Griesinger CB (2024a) Cyber security in the health and medicine sector: a study on available evidence of patient health consequences resulting from cyber incidents in healthcare settings. European Commission: Joint Research Centre. Online: <https://publications.jrc.ec.europa.eu/repository/handle/JRC138692>
629. Reina V & Griesinger CB (2024b) An ontology of health impacts and value chain enablers of digital health solutions. European Commission: Joint Research Centre. Publications Office of the European Union, Luxembourg, 2024, <https://data.europa.eu/doi/10.2760/7823548>, JRC139910
630. Reis AA et al. (2024) Future-Proofing Research Ethics—Key Revisions of the Declaration of Helsinki 2024. *JAMA*. Published online October 19, 2024. doi:10.1001/jama.2024.22254
631. Reisenzein R, Horstmann G & Schützwohl A (2017) The Cognitive-Evolutionary Model of Surprise: A Review of the Evidence. *Top Cogn Sci.* 2019 Jan;11(1):50-74. doi: 10.1111/tops.12292
632. Ricci Lara MA (2022) Addressing fairness in artificial intelligence for medical imaging. *Nat Commun* 13, 4581 Online: <https://doi.org/10.1038/s41467-022-32186-3>
633. Rivera et al. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension Online: <https://doi.org/10.1136/bmj.m3210>
634. Robbins SA (2019) A misdirected Principle with a Catch: Explicability for AI. *Minds & Machines* 29, 495–514. <https://doi.org/10.1007/s11023-019-09509-3>
635. Roch C (2024) The critical link between data quality and generative AI. Article on LinkedIn. Online: <https://www.linkedin.com/pulse/critical-link-between-data-quality-generative-ai-eric-roch-w0krcl/>
636. Roche C et al. (2021) Artificial Intelligence Ethics: An Inclusive Global Discourse? arXiv:2108.09959. Online: <https://doi.org/10.48550/arXiv.2108.09959>
637. Roche C et al. (2023) Ethics and diversity in artificial intelligence policies, strategies and initiatives. *AI Ethics* 3, 1095–1115 (2023). Online: <https://doi.org/10.1007/s43681-022-00218-9>
638. Rogers K (2025) Personalised medicine. Britannica. Online: <https://www.britannica.com/science/personalized-medicine>
639. Roit et al. (2023) Factually consistent summarization via reinforcement learning with textual entailment feedback. Online: <http://arxiv.org/abs/2306.00186>
640. Rony MKK et al. (2024) I Wonder if my Years of Training and Expertise Will be Devalued by Machines: Concerns About the Replacement of Medical Professionals by Artificial Intelligence. *SAGE Open Nursing.* 2024;10. Online: doi:10.1177/23779608241245220
641. Röösli E et al. (2022) Peeking into a black box, the fairness and generalizability of a MIMIC-III benchmarking model. *Sci Data* 9, 24. Online: <https://doi.org/10.1038/s41597-021-01110-7>
642. Rosenblatt F (1958) The perceptron: a probabilistic model for information storage and organisation in the brain. *Psychological review* 65,6: 386-408. Online: <https://www.ling.upenn.edu/courses/cogs501/Rosenblatt1958.pdf>
643. Roski J et al. How artificial intelligence is changing health and healthcare. In: Matheny et al. (eds) (2022) Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril.

- Washington, DC: National Academy of Medicine: The National Academies Press. Online: <https://doi.org/10.17226/27111>.
644. Royakkers L. et al. (2018) Societal and ethical issues of digitization. *Ethics Inf Technol* 20, 127–142. Online: <https://doi.org/10.1007/s10676-018-9452-x>
645. Rudin C (2019) Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nat Mach Intell.* 2019 May;1(5):206–215. Online: doi: 10.1038/s42256-019-0048-x. Epub 2019 May 13.
646. Ryan M & Stahl BC (2021) Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications, *Journal of Information, Communication and Ethics in Society*, Vol. 19 No. 1, pp. 61–86. Online: <https://doi.org/10.1108/JICES-12-2019-0138Ryan>

## S

647. S & P (2024) The AI governance challenge. Online: <https://www.spglobal.com/en/research-insights/special-reports/the-ai-governance-challenge>
648. Sadegh-Zadeh K (2011) The hypothetico-deductive approach. In "Philosophy of medicine", Vol.16 in Handbook of the philosophy of science. Edited by Gabbay DM, Gifford F, Thagard P and Woods J. Elsevier. Online: <https://www.sciencedirect.com/book/9780444517876/philosophy-of-medicine>
649. Sæbø S & Brovold H (2024) On the stochastics of human and artificial creativity. arXiv:2403.06996. Online: <https://doi.org/10.48550/arXiv.2403.06996>
650. Saeidi H et al. (2022) Autonomous robotic laparoscopic surgery for intestinal anastomosis. *Science robotics* 7(62) 13 pages. Online: DOI: 10.1126/scirobotics.abj2908Text
651. Saint James Aquino Y (2023) Making decisions: Bias in artificial intelligence and data-driven diagnostic tools. *Aust J Gen Pract.* Jul;52(7):439–442. Online: doi: 10.31128/AJGP-12-22-6630. PMID: 37423238.
652. Sajno E et al. (2023) Machine learning in biosignals processing for mental health: A narrative review. *Front Psychol.* 2023 Jan 13;13:1066317. Online: doi: 10.3389/fpsyg.2022.1066317
653. Salehi AW (2023) A Study of CNN and Transfer Learning in Medical Imaging: Advantages, Challenges, Future Scope. *Sustainability.* 15(7):5930. Online: <https://doi.org/10.3390/su15075930>
654. Salih AM et al. (2024) A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME. arXiv. Online: <https://arxiv.org/html/2305.02012v3>
655. Salmon WC (1989) Introduction to “Four Decades of Scientific Explanation”. Eds.:Kitcher and Salmon. Minneapolis, MN: University of Minnesota Press. Online: <https://conservancy.umn.edu/server/api/core/bitstreams/e28406fc-d9d6-412f-b4b8-cf7a800291a3/content>
656. Sambasivan N & Veeraraghavan R (2022) The Deskilling of Domain Expertise in AI Development. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 587, 1–14. Online: <https://doi.org/10.1145/3491102.3517578>
657. Samek W & Müller KR (2019) Towards explainable artificial intelligence, in: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer, Cham, Switzerland, pp. 5–22. Online: [http://dx.doi.org/10.1007/978-3-030-28954-6\\_1](http://dx.doi.org/10.1007/978-3-030-28954-6_1)
658. Samuel AL (1969) Some Moral and Technical Consequences of Automation—A Refutation. *Science* 6 Sep 1960, Vol 132, Issue 3429, pp. 741–742. Online: DOI: 10.1126/science.132.3429.741
659. Sanchez P et al. (2022) Causal machine learning for healthcare and precision medicine. *R. Soc. Open Sci.* 9.220638. Online: <http://doi.org/10.1098/rsos.220638>

660. Sarker IH (2021a) Machine learning: algorithms, real-world applications and research directions. SN Comput Sci. 2021;2(3):1-21. Online: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7983091/>
661. Sarker, I.H. (2021b) Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. SN COMPUT. SCI. 2, 420. Online: <https://doi.org/10.1007/s42979-021-00815-1>
662. Sarker IH (2022) AI-Based Modeling: Techniques, Applications and Research Issues Towards Automation, Intelligent and Smart Systems. SN COMPUT. SCI. 3, 158 (2022). Online: <https://doi.org/10.1007/s42979-022-01043-x>
663. Sarris J (2022) Disruptive innovation in psychiatry. Ann N Y Acad Sci. 2022 Jun;1512(1):5-9. Online: doi: 10.1111/nyas.14764
664. Satpute A et al (2024) Can LLMs Master Math? Investigating Large Language Models on Math Stack Exchange. arXiv. Online: <https://arxiv.org/html/2404.00344v1>
665. Saslow & Lorenz (2019) Artificial intelligence needs human rights. Stiftung Neue Verantwortung. Online: [www.stiftung-nv.de](http://www.stiftung-nv.de)
666. Scarcello F (2019) Artificial intelligence. In: Encyclopedia of Bioinformatics and Computational Biology. Vol 1: 287-293. Online: <https://www.sciencedirect.com/science/article/pii/B9780128096338203269?via%3Dihub> (Last accessed: 2024.08.22)
667. Schachner T et al. (2020) Artificial Intelligence-Based Conversational Agents for Chronic Conditions: Systematic Literature Review. J Med Internet Res. Sep 14;22(9):e20701. Online: doi: 10.2196/20701.
668. Schiff D & Borenstein J (2019) How Should Clinicians Communicate With Patients About the Roles of Artificially Intelligent Team Members? AMA J Ethics. 21(2):E138-145. Online: doi: 10.1001/ama.jethics.2019.138.
669. Schinkel M et al. (2019) Clinical applications of artificial intelligence in sepsis: A narrative review. Comput Biol Med. 2019 Dec;115:103488. doi: 10.1016/j.compbiomed.2019.103488. Epub 2019 Oct 7.
670. Schiro J et al. (2017) Usability Validation of Medical Devices: Issues in Identifying Potential Use Errors. Stud Health Technol Inform. Online: [doi:10.3233/978-1-61499-742-9-298](https://doi.org/10.3233/978-1-61499-742-9-298)
671. Schlieter H & Esswein W (2010) From Clinical Practice Guideline to Clinical Pathway –Issues of Reference Model-Based Approach. In: Camarinha-Matos, L.M., Boucher, X., Afsarmanesh, H. (eds) Collaborative Networks for a Sustainable World. PRO-VE 2010. IFIP Advances in Information and Communication Technology, vol 336. Springer, Berlin, Heidelberg. Online: [https://doi.org/10.1007/978-3-642-15961-9\\_30](https://doi.org/10.1007/978-3-642-15961-9_30)
672. Schmidhuber J (2015) Deep learning in neural networks: An overview. Neural Networks, Volume 61, Pages 85-117. Online: <https://doi.org/10.1016/j.neunet.2014.09.003>
673. Schmidt P et al. (2019) Wearable-Based Affect Recognition-A Review. Sensors (Basel). 2019 Sep 20;19(19):4079. Online: doi: 10.3390/s19194079.
674. Schulz KF et al. (2010). CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. Online: <https://doi.org/10.1136/bmj.c332>
675. Schwabe D et al. (2024) The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review. npj Digit. Med. 7, 203. Online: <https://doi.org/10.1038/s41746-024-01196-4>
676. Scottish NHS (2023) Key Definitions: Decision-Making Support Tools, Clinical Guidelines, Policies, Protocols, Procedures & Care Pathways. Online: <https://rightdecisions.scot.nhs.uk/media/2672/1-key-definitions-decision-making-support-tools.pdf>
677. SEA platform network (2024) Linking AI Principles (LAIPI). Online: <https://www.linkings-ai-principles.org/>
678. Searle JR (1980) Minds, brains, and programs. Behavioural and Brain Sciences. 1980;3(3):417-424. doi:10.1017/S0140525X00005756

679. Searle JR (1983) Intentionality: An essay in the philosophy of mind. Cambridge university press, New York, USA.
680. Searle JR (1984) Minds, brain and science. Reith lectures. Harvard university press, Cambridge (Massachusetts), USA.
681. Sedaghat S (2023) Success Through Simplicity: What Other Artificial Intelligence Applications in Medicine Should Learn from History and ChatGPT. Ann Biomed Eng 51, 2657–2658 (2023). <https://doi.org/10.1007/s10439-023-03287-x>
682. Sejnowski TJ, Koch C, Churchland PS (1988) Computational neuroscience. Science. Sep 9;241(4871):1299-306. Online: doi: 10.1126/science.3045969.
683. Selbst AD (2018) Fairness and abstraction in Sociotechnical Systems. In ACT conference on fairness, accountability, and transparency (FAT) (vol. 1, No. 1, pp. 1–17).
684. Selbst AD et al. (2019a) Fairness and Abstraction in Sociotechnical Systems. In: FAT\* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency. Pages 59 – 68. Online: <https://doi.org/10.1145/3287560.3287598>
685. Selbst AD (2019b) Negligence and AI's human users. Boston University Law Review, 1315. Online: [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3350508#](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3350508#) Web version: <https://www.private-law-theory.org/2024/01/19/andrew-selbst-negligence-and-ais-human-users-4/>
686. Selbst AD (2021) An Institutional View Of Algorithmic Impact Assessments (June 15, 2021). 35 Harvard Journal of Law & Technology 117, UCLA School of Law, Public Law Research Paper No. 21-25, Available at SSRN: <https://ssrn.com/abstract=3867634>
687. Selvaraju RR et al. (2016) Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. arXiv. Online: <https://doi.org/10.48550/arXiv.1610.02391> Associated doi: <https://link.springer.com/article/10.1007/s11263-019-01228-7> (International Journal of Computer Vision, Volume 128, pages 336–359, (2020)).
688. SEP - Stanford encyclopedia of philosophy. *bounded rationality*. Online: <https://plato.stanford.edu/entries/bounded-rationality/#TwoSchoHeur>
689. SEP - Stanford encyclopedia of philosophy: *agency*. Online: <https://plato.stanford.edu/entries/agency/> (Last accessed 13.5.2024)
690. SEP - Stanford encyclopedia of philosophy: *deontological ethics*. Online: <https://plato.stanford.edu/entries/ethics-deontological/> (Last accessed: 2024.08.19)
691. SEP - Stanford encyclopedia of philosophy: *functionalism*. Online: <https://plato.stanford.edu/entries/functionalism/>
692. SEP - Stanford encyclopedia of philosophy: *metaethics*. <https://plato.stanford.edu/entries/metaethics/>
693. SEP - Stanford Encyclopedia of philosophy: *personal autonomy*. Online: <https://plato.stanford.edu/entries/personal-autonomy/>
694. SEP - Stanford encyclopedia of philosophy: *the computational theory of mind*. Online: <https://plato.stanford.edu/entries/computational-mind/>
695. SEP - Stanford encyclopedia of philosophy: *the principle of beneficence in applied ethics*. <https://plato.stanford.edu/entries/principle-beneficence/>
696. SEP – Stanford encyclopedia of philosophy: scientific explanation. Online: <https://plato.stanford.edu/entries/scientific-explanation/#DNMode>
697. Serna P & Seoane JA (2016) Bioethical Decision Making and Argumentation. International Library of Ethics, Law, and the New Medicine 70. Springer Verlag. Online: DOI 10.1007/978-3-319-43419-3
698. Shah C (2023), Keeping Patient Data Secure in the Age of Radiology Artificial Intelligence: Cybersecurity Considerations and Future Directions, Journal of the American College of Radiology, Volume 20, Issue 9. <https://doi.org/10.1016/j.jacr.2023.06.023>

699. Shah et al., (2023) Group Fairness with Uncertainty in Sensitive Attributes. arXiv:2302.08077. Online: <https://doi.org/10.48550/arXiv.2302.08077>
700. Shah S & Robinson I (2008) Medical device technologies: who is the user? International Journal of Healthcare Technology and Management Vol. 9, No. 2: 181-197. Online via research gate: [https://www.researchgate.net/publication/49402282\\_Medical\\_device\\_technologies\\_Who\\_is\\_the\\_user](https://www.researchgate.net/publication/49402282_Medical_device_technologies_Who_is_the_user)
701. Shaw JA (2024) The Revised Declaration of Helsinki—Considerations for the Future of Artificial Intelligence in Health and Medical Research. JAMA. Published online October 19, 2024. doi:10.1001/jama.2024.22074
702. Shelf (2024) Fairness Metrics in AI—Your Step-by-Step Guide to Equitable Systems. Online: <https://shelf.io/blog/fairness-metrics-in-ai/#:~:text=Fairness%20metrics%20are%20quantitative%20measures,or%20amplifying%20biases%20and%20inequalities>. (Last accessed: 2024.09.05)
703. Shen T (2023) Large Language Model Alignment: A Survey. arXiv. Online: arXiv:2309.15025
704. Shetty R (2025) LLM Evaluation: 15 Metrics You Need to Know. Online: <https://arya.ai/blog/llm-evaluation-metrics>
705. Shi Q et al. (2025) Hybrid neural networks for continual learning inspired by corticohippocampal circuits. Nat Commun 16, 1272. Online: <https://doi.org/10.1038/s41467-025-56405-9>
706. Shin & Kim (2024) National Assembly passes the AI Basic Act. Online: <https://www.shinkim.com/eng/media/newsletter/2667>
707. Shneiderman B (2020) Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems. ACM Transactions on Interactive Intelligent Systems, 10(4), 1–31. Online: <https://dl.acm.org/doi/10.1145/3419764>
708. Shumway DO, Hartman HJ. (2024) Medical malpractice liability in large language model artificial intelligence: legal review and policy recommendations. J Osteopath Med. Jan 31;124(7):287-290. doi: 10.1515/jom-2023-0229.
709. Silcox C et al. (2024) The potential for artificial intelligence to transform healthcare: perspectives from international health leaders. NPJ Digit Med. Apr 9;7(1):88. Online: doi: 10.1038/s41746-024-01097-6.
710. Simoncini A & Longo E (2021) Fundamental Rights and the Rule of Law in the Algorithmic Society. In: Micklitz H-W, Pollicino O, Reichman A, Simoncini A, Sartor G, De Gregorio G, eds. Constitutional Challenges in the Algorithmic Society. Cambridge University Press; 27-41. Online: <https://www.cambridge.org/core/books/constitutional-challenges-in-the-algorithmic-society/fundamental-rights-and-the-rule-of-law-in-the-algorithmic-society/478713311AA2EF9BD742B2146726911E>
711. Simonyan K, Vedaldi A, Zisserman A (2013) Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv:1312.6034. Online: <https://doi.org/10.48550/arXiv.1312.6034>
712. Sinosoglou I et al. (2023) Post-Processing Fairness Evaluation of Federated Models: An Unsupervised Approach in Healthcare. IEEE/ACM Trans Comput Biol Bioinform. 2023 Jul-Aug; 20(4):2518-2529. doi: 10.1109/TCBB.2023.3269767. Epub 2023 Aug 9. PMID: 37097792.
713. Sinosoglou I et al. (2023) Post-Processing Fairness Evaluation of Federated Models: An Unsupervised Approach in Healthcare. IEEE/ACM Trans Comput Biol Bioinform. 2023 Jul-Aug;20(4):2518-2529. doi: 10.1109/TCBB.2023.3269767. Epub 2023 Aug 9. PMID: 37097792.
714. Sloane M & Moss E (2023) Assessing the Assessment: Comparing Algorithmic Impact Assessments and AI Audits In review for edited volume for Oxford University Press. June 20, 2023. ONLINE at SSRN: <https://ssrn.com/abstract=4486259> or <http://dx.doi.org/10.2139/ssrn.4486259>
715. Smith E et al. (2021) Affective Computing for Late-Life Mood and Cognitive Disorders. Front Psychiatry. 2021 Dec 23;12:782183. Online: doi: 10.3389/fpsyg.2021.782183.

716. Smith C (2025) Expert comment: calling for global AI legislation at the AI Action Summit. University of Salford/Manchester. Online: <https://www.salford.ac.uk/news/expert-comment-calling-for-global-ai-legislation-at-the-ai-action-summit>
717. Soares N et al. (2015) Corrigibility. In AAAI Workshops: Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, January 25–26, 2015. AAAI Publications. Online: <https://intelligence.org/files/Corrigibility.pdf> (Last accessed: 2024.08.12)
718. Soliman A et al. (2023) The Price of Explainability in Machine Learning Models for 100-Day Readmission Prediction in Heart Failure: Retrospective, Comparative, Machine Learning Study. *J Med Internet Res.* Oct 27;25:e46934. Online: doi: 10.2196/46934
719. Song W et al. (2020) Human factors risk assessment: An integrated method for improving safety in clinical use of medical devices, *Applied Soft Computing*, Volume 86, 105918, ISSN 1568-4946, <https://doi.org/10.1016/j.asoc.2019.105918>.
720. Sornette D et al. (2007) Algorithm for model validation: theory and applications. *Proc Natl Acad Sci USA.* 2007 Apr 17;104(16):6562–7. Online: doi: 10.1073/pnas.0611677104
721. Sounderajah V et al. (2020) Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group. Online: [doi.org/10.1038/s41591-020-0941-1](https://doi.org/10.1038/s41591-020-0941-1)
722. Sounderajah V et al. (2021a) Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. Online: [doi.org/10.1136/bmjopen-2020-047709](https://doi.org/10.1136/bmjopen-2020-047709)
723. Sounderajah, V et al. (2021b) A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. *Nat Med* 27, 1663–1665. Online: <https://doi.org/10.1038/s41591-021-01517-0>
724. Spirtes P, Glymour C, & Scheines R (2001) *Causation, prediction, and search*. MIT press. Online: <https://doi.org/10.7551/mitpress/1754.001.0001>
725. Srivastava N et al., (2014) Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15: 1929–1958. Online: <https://jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf>
726. Stahl BC et al. (2023) A systematic review of artificial intelligence impact assessments. *Artif Intell Rev* 56, 12799–12831 (2023). <https://doi.org/10.1007/s10462-023-10420-8>. (Stahl BC et al. have listed 38 AI-specific impact assessment frameworks in a publicly accessible Zotero group library. Online: [https://www.zotero.org/groups/4042832/ai\\_impact\\_assessments/library](https://www.zotero.org/groups/4042832/ai_impact_assessments/library))
727. Stevens AF & Stetson P (2023) Theory of trust and acceptance of artificial intelligence technology (TrAAIT): An instrument to assess clinician trust and acceptance of artificial intelligence. *J Biomed Inform.* 2023 Dec;148:104550. Online: doi: 10.1016/j.jbi.2023.104550. Epub 2023 Nov 20.
728. Straw I (2020) The automation of bias in medical Artificial Intelligence (AI): Decoding the past to create a better future. *Artif Intell Med.* Nov;110:101965. Online: doi: 10.1016/j.artmed.2020.101965. Epub 2020 Oct 6.
729. Sujan M (2023) Validation framework for the use of AI in healthcare: overview of the new British standard BS30440. *BMJ Health Care Inform.* 2023 Jun;30(1):e100749. doi: 10.1136/bmjhci-2023-100749. N.B. the relevant standard is behind a paywall.
730. Sun Y et al. (2024) AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content. *Humanit Soc Sci Commun* 11, 1278. Online: <https://doi.org/10.1057/s41599-024-03811-x>
731. Sun YJ et al. (2019) Experience-dependent structural plasticity at pre- and postsynaptic sites of layer 2/3 cells in developing visual cortex. *Proc Natl Acad Sci U S A.* Oct 22;116(43):21812–21820. Online: doi: 10.1073/pnas.1914661116.
732. Suresh H & Guttag JV (2021) A framework for understanding sources of harm throughout the machine learning life cycle. *arXiv:1901.10002*. Online: <https://doi.org/10.48550/arXiv.1901.10002> (Last accessed: 2024.08.20)

## T

733. Talaei Khoei T et al. (2023) Deep learning: systematic review, models, challenges, and research directions. *Neural Comput & Applic* 35, 23103–23124 (2023). Online: <https://doi.org/10.1007/s00521-023-08957-4>
734. Tamò-Larrieux A et al. (2024) Regulating for trust: Can law establish trust in artificial intelligence? *Regulation & Governance* (2024) 18, 780–801. Online: doi:10.1111/rego.12568
735. Tanaka S et al (2020) Development and Reorganisation of Orientation Representation in the Cat Visual Cortex: Experience-Dependent Synaptic Rewiring in Early Life. *Front Neuroinform.* Aug 20;14:41. Online: doi: 10.3389/fninf.2020.00041
736. Tarabanis C et al. (2023) Development of an AI-Driven QT Correction Algorithm for Patients in Atrial Fibrillation. *JACC Clin Electrophysiol.* Feb;9(2):246–254. Online: doi: 10.1016/j.jacep.2022.09.021
737. TechTarget (2024) AI governance. Online: <https://www.techtarget.com/searchenterpriseai/definition/AI-governance>
738. Tejani AS et al. (2024) Understanding and Mitigating Bias in Imaging Artificial Intelligence. *Radiographics.* May;44(5):e230067. Online: doi: 10.1148/rg.230067.
739. Tejani AS, et al. (2024b). CLAIM 2024 Update Panel. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update. *Radiol Artif Intell.* 2024 Jul;6(4):e240300. <https://doi.org/10.1148/ryai.240300>
740. Terven J et al. (2023) Loss functions and metrics in deep learning. arXiv:2307.02694. Online: <https://doi.org/10.48550/arXiv.2307.02694>
741. Testolin A (2023) Can neural networks do arithmetic? A survey on the elementary numerical skills of state-of-the-art deep learning models. Online: <https://doi.org/10.3390/app14020744>
742. Thapa C et al., (2020) SplitFed: When Federated Learning Meets Split Learning. arXiv:2004.12088. Online: <https://doi.org/10.48550/arXiv.2004.12088>
743. The “Joint Commission” (2024) website: Implicit bias in healthcare. Online: <https://www.joint-commission.org/resources/news-and-multimedia/newsletters/newsletters/quick-safety/quick-safety-issue-23-implicit-bias-in-health-care/implicit-bias-in-health-care/> (Last accessed: 2024.09.07)
744. The European Institute of Innovation and Technology (EIT) (2021), Creation of a taxonomy for the European AI ecosystem. Online: <https://eit.europa.eu/library/creation-taxonomy-european-ai-ecosystem> .
745. Thirunavukarasu AJ et al. (2023) Large language models in medicine. *Nat Med* 29, 1930–1940 (2023). <https://doi.org/10.1038/s41591-023-02448-8>
746. Tocchetti A et al. (2022) A.I. Robustness: a Human-Centered Perspective on Technological Challenges and Opportunities. Online: <https://doi.org/10.48550/arXiv.2210.08906>
747. Tran A, Luong T, Huynh V (2024) A comprehensive survey and taxonomy on privacy-preserving deep learning. *Neurocomputing*, Volume 576, 127345, ISSN 0925-2312. Online: <https://doi.org/10.1016/j.neucom.2024.127345>
748. Tripp CE et al. (2024) Measuring the Energy Consumption and Efficiency of Deep Neural Networks: An Empirical Analysis and Design Recommendations. arXiv:2403.08151v1 [cs.LG] 13 Mar 2024
749. Tschantz MC 2022. What is Proxy Discrimination? In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Association for Computing Machinery, New York, NY, USA, 1993–2003. Oline: <https://doi.org/10.1145/3531146.3533242>
750. Tsopra (2021) A framework for validating AI in precision medicine: considerations from the European ITFoC consortium. *BMC Med Inform Decis Mak.* 2021. <https://doi.org/10.1186/s12911-021-01634-3>

751. Tsvetkova et al. (2017) Understanding Human-Machine Networks: A Cross-Disciplinary Survey. ACM Computing Surveys Volume 50 Issue 1 Article No.: 12pp 1–35. Online: <https://doi.org/10.1145/3039868>
752. Tu T et al. (2023) Towards Generalist Biomedical AI. Online: arXiv:2307.14334 or <https://doi.org/10.48550/arXiv.2307.14334>
753. Turing (website) (2025) What is the necessity of bias in neural networks? Online: <https://www.turing.com/kb/necessity-of-bias-in-neural-networks#what-is-bias-in-a-neural-network?>
754. Tyler J (2024) AI and Semantic Search: Crafting Content for Conceptual Relevance. Online: <https://medium.com/@jamestryrh/ai-and-semantic-search-crafting-content-for-conceptual-relevance-4e6670533326>

## U

755. UK Department for environment, food and rural affairs (2017) HSAC paper on definition of lifecycle/value chain in relation to nanomaterials and other manufactured substances. Online: <https://assets.publishing.service.gov.uk/media/5a80e9ceed915d74e623127e/hsac-paper-definition-lifecycle-value-chain.pdf>
756. UK Competition and Markets Authority (2023). AI Foundation Models: Initial Report. Online: [https://assets.publishing.service.gov.uk/media/65081d3aa41cc300145612c0/Full\\_report\\_.pdf](https://assets.publishing.service.gov.uk/media/65081d3aa41cc300145612c0/Full_report_.pdf)
757. UK National Health Service (NHS) 2024) Chapter 3.2 evaluation and validation. In: Understanding healthcare workers' confidence in artificial intelligence (AI) (website training resource). Online: <https://digital-transformation.hee.nhs.uk/building-a-digital-workforce/dart-ed/horizon-scanning/understanding-healthcare-workers-confidence-in-ai>
758. UK government (2023) The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023. Online: <https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023>
759. UK government (2025) Department for science, innovation and technology. Online: <https://www.gov.uk/government/publications/international-ai-safety-report-2025/international-ai-safety-report-2025>

## United Nations

760. United Nations (UN) (1948) Universal Declaration of Human Rights. Online: <https://www.un.org/en/about-us/universal-declaration-of-human-rights>
761. United Nations (UN) (1966) International covenant on civil and political rights. Online: <https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights>
762. United Nations (UN) (1987) Brundtland commission report. Online: <http://www.un-documents.net/our-common-future.pdf>
763. United Nations (UN) Development Group (2017) Data privacy, ethics and protection. Guidance note on big data for achievement of the 2030 agenda. Online: [https://unsdg.un.org/sites/default/files/UNDG\\_BigData\\_final\\_web.pdf](https://unsdg.un.org/sites/default/files/UNDG_BigData_final_web.pdf)
764. United Nations (UN) Interregional Crime and Justice Research Institute - Center for Artificial Intelligence and Robotics (2020) "Special collection on artificial intelligence" (chapter 5: AI in healthcare: risk assessment and criminal law). Online: <https://unicri.it/News/Artificial%20Intelligence%20Collection>
765. United Nations (UN) – UNESCO (2021) Recommendation on the ethics of artificial intelligence. Online: <https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence>

766. United Nations (UN) (2021) – UNESCO Recommendations on the ethics of artificial intelligence. Online: <https://unesdoc.unesco.org/ark:/48223/pf0000381137>
767. United Nations (UN) (2024a) High-level committee on programmes. Inter-Agency Working Group on Artificial Intelligence (IAWG-AI): United Nations System White Paper on AI Governance: An analysis of the UN system's institutional models, functions, and existing international normative frameworks applicable to AI governance. Online: <https://unsceb.org/sites/default/files/2024-05/United%20Nations%20System%20White%20Paper%20on%20AI%20Governance.pdf> (Last accessed: 2024.09.30)
768. United Nations (UN) – UNESCO (2024b) Ethics of Artificial Intelligence. Online: <https://www.unesco.org/en/artificial-intelligence/recommendation-ethics>
769. United Nations (UN) (2024c) United Nations System White Paper on AI Governance. Online: <https://unsceb.org/united-nations-system-white-paper-ai-governance>
770. United Nations (UN) – Inclusive policy lab (2024d). Addressing digital colonialism: A path to equitable data governance. Online: <https://en.unesco.org/inclusivepolicylab/analytics/addressing-digital-colonialism-path-equitable-data-governance>
771. United Nations (UN) (2024e) Sustainability. Online: <https://www.un.org/en/academic-impact/sustainability>
772. United Nations (UN) (2024f) UN sustainable development goals. Online: <https://www.un.org/sustainabledevelopment/sustainable-development-goals/>
773. United States - Department of Health and Human Services (2021) Trustworthy AI (TAI) playbook. Online: <https://www.hhs.gov/sites/default/files/hhs-trustworthy-ai-playbook.pdf> (Last accessed: 2024.09.20).
774. United States government, White House (2023) Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. Online: <https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/>
775. United States government, White House (2025) Executive Order on initial rescissions of harmful executive orders and actions.
776. United States - Department of Commerce - National institute of standards and technology (NIST) (2023) Artificial intelligence risk management framework. Document AI 100-1. Online: <https://www.nist.gov/itl/ai-risk-management-framework> Publicly available, no paywall. Readers should take note that the document does not contain references to scientific or other publications.
777. United States - Department of Commerce / National institute of standards and technology (NIST) (2014) Guidelines for media sanitization. NIST special publication 800-88. Online: <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-88r1.pdf>
778. Upton G & Cook I (2014) Oxford dictionary of statistics. Online: <https://www.oxfordreference.com/display/10.1093/acref/9780199679188.001.0001/acref-9780199679188>

## US FDA

779. US Food and Drug Administration (US FDA), Center for Devices and Radiological Health - CDRH (2002). General Principles of Software Validation; Final Guidance for Industry and FDA Staff. Online: <https://www.fda.gov/media/73141/download>
780. US\_Food and Drug Administration (US FDA) (2016) Applying Human Factors and Usability Engineering to Medical Devices. Online: <https://www.fda.gov/media/80481/download>
781. US Food and Drug Administration (US FDA), Health Canada, UK MHRA (2021) Good Machine Learning Practice for Medical Device Development. Online: <https://www.fda.gov/media/153486/download>
782. US Food and Drug Administration (US FDA) (2024a) FDA Digital Health and Artificial Intelligence Glossary – Educational Resource. Online: <https://www.fda.gov/science-research/artificial->

- intelligence-and-medical-products/fda-digital-health-and-artificial-intelligence-glossary-educational-resource#:~:text=Artificial%20Intelligence%20(AI),influencing%20real%20or%20virtual%20environments
783. US Food and Drug Administration (US FDA) (2024b) Artificial Intelligence and Medical Products: How CBER, CDER, CDRH, and OCP are Working Together. Online: <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device>
784. US department of health and human services (1991) Federal Policy for the Protection of Human Subjects ('Common Rule'). Online: <https://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html>
785. United States government - Centers of Excellence – Artificial Intelligence. Subsite: Understanding and managing the AI lifecycle. Online <https://coe.gsa.gov/coe/ai-guide-for-government/understanding-managing-ai-lifecycle/>. Accessed 2024.03.21.
786. United States National Institutes of Health (2024) Website: Ethics in clinical research. Online: <https://www.cc.nih.gov/recruit/ethics>
787. Unitrends (2024) The CIA Triad and Its Importance in Data Security. Online: <https://www.unitrends.com/blog/cia-triad-confidentiality-integrity-availability>

## V

788. Valois P, Niinuma K & Fukui K (2023) Occlusion Sensitivity Analysis with Augmentation Subspace Perturbation in Deep Feature Space. arXiv. Online: <https://doi.org/10.48550/arXiv.2311.15022>
789. Van Calster B & Vickers AJ (2015) Calibration of risk prediction models: impact on decision-analytic performance. Med Decis Mak. 2015;35:162–9. Online: DOI: 10.1177/0272989X14547233
790. Van Calster B et al. (2016) A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol. 74:167–76. Online: [https://www.jclinepi.com/article/S0895-4356\(15\)00581-8/fulltext](https://www.jclinepi.com/article/S0895-4356(15)00581-8/fulltext)
791. Van Calster B et al. (2019) Calibration: the Achilles heel of predictive analytics. BMC Med 17, 230 (2019). Online: <https://doi.org/10.1186/s12916-019-1466-7>
792. Van Daalen OL (2023) The right to encryption: Privacy as preventing unlawful access, Computer Law & Security Review, Volume 49. Online: <https://doi.org/10.1016/j.clsr.2023.105804>
793. van der Velden BHM (2022) Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med Image Anal. Jul;79:102470. Online: doi: 10.1016/j.media.2022.102470.
794. van Smeden et al. (2021) Clinical prediction models: diagnosis versus prognosis. J Clin Epidemiol. 2021 Apr;132:142–145. Online: doi: 10.1016/j.jclinepi.2021.01.009
795. Varela F (1992) Whence perceptual meaning. A Cartography of Current Ideas. In: Varela, F.J., Dupuy, JP. (eds) Understanding Origins. Boston Studies in the Philosophy and History of Science, Vol 130. Springer, Dordrecht. Online: [https://doi.org/10.1007/978-94-015-8054-0\\_13](https://doi.org/10.1007/978-94-015-8054-0_13)
796. Varkey B (2021) Principles of Clinical Ethics and Their Application to Practice. Med Princ Pract 2021;30:17–28 Online: DOI: 10.1159/000509119
797. Vasey et al. (2022) Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Online: <https://doi.org/10.1136/bmj-2022-070904>
798. Vaswani et al., (2017) Attention is all you need. arXiv. Online: <https://doi.org/10.48550/arXiv.1706.03762>

799. Veale M & Binns R (2017) Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, 4(2), 1–17. Online: <https://doi.org/10.1177/2053951717743530>
800. Veit-Haibach P & Herrmann K eds. (2022) Artificial Intelligence/Machine Learning in Nuclear Medicine and Hybrid Imaging. Online: <https://link.springer.com/book/10.1007/978-3-031-00119-2>
801. Vela MB (2022) Eliminating Explicit and Implicit Biases in Health Care: Evidence and Research Needs. *Annu Rev Public Health*. Apr 5;43:477–501. Online: doi: 10.1146/annurev-publhealth-052620-103528. Epub 2022 Jan 12.
802. Verghese A et al. (2018) What This Computer Needs Is a Physician: Humanism and Artificial Intelligence. *JAMA*. 2018;319(1):19–20. Online: doi:10.1001/jama.2017.19198 (Last accessed: 2024.08.12)
803. Veselovsky v et al. (2023) Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks. Online: <https://doi.org/10.48550/arXiv.2306.07899>
804. Vilone G & Longo L (2021) Notions of explainability and evaluation approaches for explainable artificial intelligence, *Information Fusion*, Volume 76, Pages 89–106, Online: <https://doi.org/10.1016/j.inffus.2021.05.009>
805. Vicente, AM (2020) How personalised medicine will transform healthcare by 2030: the ICPMed vision. *J Transl Med* 18, 180 Online: <https://doi.org/10.1186/s12967-020-02316-w>
806. Vokinger KN et al. (2021) Mitigating bias in machine learning for medicine. *Commun Med* 1, 25. Online: <https://doi.org/10.1038/s43856-021-00028-w>
807. Vorisek CN et al. (2023) Artificial Intelligence Bias in Health Care: Web-Based Survey. *J Med Internet Res*. Jun 22;25:e41089. Online: doi: 10.2196/41089.
808. Vyas DA et al. (2020) Hidden in Plain Sight - Reconsidering the Use of Race Correction in Clinical Algorithms. *N Engl J Med*. Aug 27;383(9):874–882. Online: doi: 10.1056/NEJMms2004740 . Epub 2020 Jun 17.

## W

809. Wagner JK et al. (2024) AI Governance: A Challenge for Public Health. *JMIR Public Health Surveill*. Sep 30;10:e58358. Online: doi: 10.2196/58358
810. Wallach WC, Allen C, Smit I (2008) Machine morality: Bottom-up and top-down approaches for modelling human moral faculties. *AI & Society* 22(4): 565–582. Online: <https://link.springer.com/article/10.1007/s00146-007-0099-0>
811. Wang J et al. (2024) Deep learning on medical image analysis. *CAAI transactions on intelligence technology*. Online: <https://doi.org/10.1049/cit2.12356>
812. Wang L et al. (2023) A Comprehensive Survey of Continual Learning: Theory, Method and Application. Online: arXiv:2302.00487 or <https://doi.org/10.48550/arXiv.2302.00487>
813. Wang RY & Strong DM (1996) What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, Vol. 12, No. 4 (Spring, 1996), pp. 5-33. Online: <http://www.jstor.org/stable/40398176>
814. Wang Z (2016) Dueling Network Architectures for Deep Reinforcement Learning. arXiv:1511.06581 or arXiv:1511.06581v3. Online: <https://doi.org/10.48550/arXiv.1511.06581>
815. Wang X et al. (2021) Deep virtual adversarial self-training with consistency regularization for semi-supervised medical image classification. *Med Image Anal*. May;70:102010. doi: 10.1016/j.media.2021.102010.
816. Wang XJ et al. (2020) Computational neuroscience: a frontier of the 21st century. *Natl Sci Rev*. 2020 Jun 12;7(9):1418–1422. doi: 10.1093/nsr/nwaa129

817. Wasson JH (1985) Clinical prediction rules. Applications and methodological standards. *N Engl J Med.* Sep 26;313(13):793-9. Online: doi: 10.1056/NEJM198509263131306
818. Wegwarth O, Gaissmaier W, Gigerenzer G. (2009) Smart strategies for doctors and doctors-in-training: heuristics in medicine. *Med Educ.* Online: doi:10.1111/j.1365-2923.2009.03359.x.
819. Weld DS & Bansal G (2019) The challenge of crafting intelligible artificial intelligence. *Communications of the ACM*, June 2019, Vol. 62 No. 6, Pages 70-79. Online: <https://dl.acm.org/doi/pdf/10.1145/3282486> (Last accessed: 2024.08.16) An earlier version of this paper entitled "Intelligible artificial intelligence" was published in 2018. Online at the National Science Foundation: <https://par.nsf.gov/servlets/purl/10075101> (Last accessed: 2024.08.16).
820. Welte, D et al. (2023) FAIR in action - a flexible framework to guide FAIRification. *Sci Data* 10, 291. Online: <https://doi.org/10.1038/s41597-023-02167-2>
821. Westbury CF (2010) Bayes' rule for clinicians: an introduction. *Front Psychol.* Nov 16;1:192. Online: doi: 10.3389/fpsyg.2010.00192
822. Whelehan DF et al. (2020) Medicine and heuristics: cognitive biases and medical decision-making. *Ir J Med Sci.* 2020 Nov;189(4):1477-1484. Online: doi: 10.1007/s11845-020-02235-1
823. Whittaker M (2018) AI now report 2018 (pp. 1–62). Online: [https://ainowinstitute.org/wp-content/uploads/2023/04/AI\\_Now\\_2018\\_Report.pdf](https://ainowinstitute.org/wp-content/uploads/2023/04/AI_Now_2018_Report.pdf)
824. Whittlestone et al. (2019) The role and limits of principles in AI ethics: towards a focus on tensions. *AIES '19: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 195-200. Online: <https://doi.org/10.1145/3306618.3314289>
825. Wiener N (1948) Cybernetics or Control and Communication in the Animal and the Machine. MIT press. Online: <https://doi.org/10.7551/mitpress/11810.001.0001>
826. Wiener N (1960) Some Moral and Technical Consequences of Automation. *Science*. 6 May 1960. Vol 131, Issue 3410, pp. 1355-1358. Online: DOI: 10.1126/science.131.3410.1355
827. Widjaja JT (2024) Successful AI Ethics & Governance at Scale: Bridging The Interpretation Gap. Medium. Online: <https://medium.com/data-science/successful-ai-ethics-governance-at-scale-bridging-the-interpretation-gap-a8249b547e62>
828. Wikipedia (2024) Gradient descent. Online: [https://en.wikipedia.org/wiki/Gradient\\_descent](https://en.wikipedia.org/wiki/Gradient_descent)
829. Wilkinson MD (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data.* Mar 15;3:160018. doi: 10.1038/sdata.2016.18. Erratum in: *Sci Data.* 2019 Mar 19;6(1):6. doi: 10.1038/s41597-019-0009-6.

### *World Economic Forum*

830. World Economic Forum (2023) The AI divide between the global north and the global south. Online: <https://www.weforum.org/agenda/2023/01/davos23-ai-divide-global-north-global-south/>
831. World Economic Forum (2024a) 4 ways public-private partnerships can bridge the AI opportunity gap. Online: <https://www.weforum.org/stories/2024/01/public-private-partnerships-ai-re-skilling/>
832. World Economic Forum (2024b) AI governance alliance. Online: <https://initiatives.weforum.org/ai-governance-alliance/home>

### *World Health Organisation (WHO)*

833. World Health Organisation (WHO) (1946; with several amendments since) Constitution of the World Health Organisation. Online: <https://www.who.int/about/governance/constitution>
834. World Health Organisation (WHO) (2003) Medical device regulations – global overview and guiding principles. Online: <https://iris.who.int/handle/10665/42744>

835. World Health Organisation (WHO) (2009) Conceptual framework for the international classification for patient safety. Online: [https://iris.who.int/bitstream/handle/10665/70882/WHO\\_IER\\_PSP\\_2010.2\\_eng.pdf?sequence=1](https://iris.who.int/bitstream/handle/10665/70882/WHO_IER_PSP_2010.2_eng.pdf?sequence=1)
836. World Health Organisation (WHO) (2018) Declaration of Astana. Global Conference on Primary Health Care, Astana, 25–26 October 2018. Geneva. Online: <https://www.who.int/docs/default-source/primary-health/declaration/gcphc-declaration.pdf> (Last accessed 2024.09.02)
837. World Health Organisation (WHO) (2021a) Ethics and governance of artificial intelligence for health. WHO guidance. Online: <https://www.who.int/publications/i/item/9789240029200>
838. World Health Organisation (WHO) (2021b) Generating Evidence for Artificial Intelligence Based Medical Devices: A Framework for Training, Validation and Evaluation. Online: <https://www.who.int/publications/i/item/9789240038462>
839. World Health Organisation (WHO) (2021c) Global strategy on digital health 2020–2025. Online: <https://iris.who.int/bitstream/handle/10665/344249/9789240020924-eng.pdf?sequence=1> (Last accessed 2024.09.02).
840. World Health Organisation (WHO) (2021d) Health promotion glossary of terms 2021. Online: <https://www.who.int/publications/i/item/9789240038349>
841. World Health Organisation (WHO) (2021e) Global report on ageism. Online: <https://www.who.int/publications/i/item/9789240016866>
842. World Health Organisation (WHO) (2022) Ageism in artificial intelligence for health. Online: <https://www.who.int/publications/i/item/9789240040793>
843. World Health Organisation (WHO) (2023) Regulatory considerations on artificial intelligence for health. Online: <https://www.who.int/publications/i/item/9789240078871>
844. World Health Organisation (WHO) (2024a) Ethics and governance of artificial intelligence for health. Guidance on large multi-modal models. Online: <https://www.who.int/publications/i/item/9789240084759>
845. World Health Organisation (WHO) (2024b) Social determinants of health. Online: [https://www.who.int/health-topics/social-determinants-of-health#tab=tab\\_1](https://www.who.int/health-topics/social-determinants-of-health#tab=tab_1)
846. World health organisation (WHO) (2024c) Website on health technology assessment. Online: [https://www.who.int/health-topics/health-technology-assessment#tab=tab\\_1](https://www.who.int/health-topics/health-technology-assessment#tab=tab_1) (Last accessed: 2024.08.20)
847. World health organisation (WHO) (2024d) Role of social protection in reducing the burden of public health and social measures during the COVID-19 pandemic. Online: <https://www.who.int/publications/i/item/9789240100831>
848. World Medical Association (WMA) (1981) Declaration of Lisbon on the rights of patients. Online: <https://www.wma.net/wp-content/uploads/2005/09/Declaration-of-Lisbon-1981.pdf>
849. World Medical Association (WMA) (2024). World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Participants. JAMA. Online: doi:10.1001/jama.2024.21972
850. Wornow M et al. (2023) The shaky foundations of large language models and foundation models for electronic health records. npj Digit. Med. 6, 135 (2023). <https://doi.org/10.1038/s41746-023-00879-8>
851. Wright D (2011) A framework for the ethical impact assessment of information technology. Ethics Inf Technol 13:199–226. Online: <https://doi.org/10.1007/s10676-010-9242-6>
852. Wright D & Friedewald M (2013) Integrating privacy and ethical impact assessments, *Science and Public Policy*, Volume 40, Issue 6, December 2013, Pages 755–766. Online: <https://doi.org/10.1093/scipol/sct083>
853. Wu E et al. (2021) How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nat Med. Apr;27(4):582–584. Online: doi: 10.1038/s41591-021-01312-x

## X

854. Xenonstack (2023) How generative AI can improve data quality. Online: <https://medium.com/xenonstack-ai/how-generative-ai-can-improve-data-quality-d6e321261f32> (last accessed 2025.01.17)
855. Xu Z et al. (2022) Addressing Fairness Issues in Deep Learning-Based Medical Image Analysis: A Systematic Review. arXiv:2209.13177. Online: <https://doi.org/10.48550/arXiv.2209.13177>

## Y

856. Yam K (2025) Shortcut learning – the coming disaster for AI? Online: <https://dps.de/en/news/shortcut-learning-the-coming-disaster-for-ai/>
857. Yang B et al. (2021) QUADAS-C: A Tool for Assessing Risk of Bias in Comparative Diagnostic Accuracy Studies. Ann Intern Med. Nov;174(11):1592-1599. Online: doi: 10.7326/M21-2234
858. Yang X et al. (2022) A large language model for electronic health records. npj Digit. Med. 5, 194 (2022). <https://doi.org/10.1038/s41746-022-00742-2>
859. Yang Y et al. (2024) A survey of recent methods for addressing AI fairness and bias in biomedicine. J Biomed Inform. Jun; 154:104646. Online: doi: 10.1016/j.jbi.2024.104646.
860. Yu S et al. (2017) Surrogate-assisted feature extraction for high-throughput phenotyping. J Am Med Inform Assoc. 2017 Apr 1;24(e1):e143-e149. Online: doi: 10.1093/jamia/ocw135
861. Yu T & Zhu H (2020) Hyper-parameter optimisation: a review of algorithms and applications. arXiv:2003.05689. Online: <https://doi.org/10.48550/arXiv.2003.05689>
862. Xun Y et al (2023) Protocols for clinical practice guidelines. J Evid Based Med. 2023 Mar;16(1):3-9. Online: doi: 10.1111/jebm.12502

## Z

863. Zanger-Tishler M et al. (2024) Risk scores, label bias, and everything but the kitchen sink. Sci Adv. Mar 29;10(13):eadi8411. doi: 10.1126/sciadv.adl8411. Epub 2024 Mar 29.
864. Zeng Y LE & Huangfu C (2019) Linking artificial intelligence principles. In: Proceedings of the AAAI Workshop on Artificial Intelligence Safety, Honolulu, Hawaii, 2019. Aachen: CEUR Workshop Proceedings; 2019. Online: <https://arxiv.org/ftp/arxiv/papers/1812/1812.04814.pdf> (Last accessed 2024.08.05).
865. Zhang F et al. (2023) Unified fair federated learning for digital healthcare. Patterns (N Y). Dec 28;5(1):100907. doi: 10.1016/j.patter.2023.100907. PMID: 38264718; PMCID: PMC10801255.
866. Zhang K et al. (2025) Artificial intelligence in drug development. Nat Med. Jan;31(1):45-59. Online: doi: 10.1038/s41591-024-03434-4. Epub 2025 Jan 20.
867. Zhang BH et al., (2018) Mitigating Unwanted Biases with Adversarial Learning. arXiv:1801.07593. Online: <https://doi.org/10.48550/arXiv.1801.07593>
868. Zhang, M et al. (2024) Dual-attention transformer-based hybrid network for multi-modal medical image segmentation. Sci Rep 14, 25704. Online: <https://doi.org/10.1038/s41598-024-76234-y>
869. Zhang Y et al. (2022) Applications of Explainable Artificial Intelligence in Diagnosis and Surgery. Diagnostics (Basel) 12(2):237. Online: doi: 10.3390/diagnostics12020237
870. Zhao G et al. (2022) Diagnose Like a Radiologist: Hybrid Neuro-Probabilistic Reasoning for Attribute-Based Medical Image Diagnosis. IEEE Trans Pattern Anal Mach Intell. Nov;44(11):7400-7416. Online: doi: 10.1109/TPAMI.2021.3130759
871. Zheng C et al. (2023) A Slowly Progressing Neuroendocrine Tumor. ACG Case Rep J. 2023 Sep 21;10(9):e01147. Online: doi: 10.14309/crj.00000000000001147.

872. Zhou et al. (2023) Predictable artificial intelligence. Online: <https://arxiv.org/ftp/arxiv/papers/2310/2310.06167.pdf> (Last accessed: 2024.08.16)
873. Zhou Y et al., (2024) A Survey on Data Quality Dimensions and Tools for Machine Learning. arXiv:2406.19614. Online: <https://doi.org/10.48550/arXiv.2406.19614>
874. Zhu D & Wang D (2023) Transformers and their application to medical image processing: A review. Journal of Radiation Research and Applied Sciences. Volume 16, Issue 4. Online: <https://doi.org/10.1016/j.jrras.2023.100680>.
875. Ziegler DM et al. (2019) Fine-tuning language models from human preferences. arXiv. Online: arXiv:1909.08593.
876. Zou J & Schiebinger L (2021) Ensuring that biomedical AI benefits diverse populations. EBio-Medicine 67 (2021) 103358. Online: [https://www.thelancet.com/journals/ebiom/article/PIIS2352-3964\(21\)00151-1/fulltext](https://www.thelancet.com/journals/ebiom/article/PIIS2352-3964(21)00151-1/fulltext)
877. Zouari S (2021) All You Need To Understanding Activation Function In Neural Networks. Medium. Online: <https://salmenzouari.medium.com/all-you-need-to-understanding-activation-function-in-neural-networks-41a00717f774>
878. Zowghi D, Bano M (2024) AI for all: Diversity and Inclusion in AI. AI Ethics. Vol. 4:873-76. Online: <https://doi.org/10.1007/s43681-024-00485-8>
879. Zvyagin M et al. (2022) GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics. The International Journal of High Performance Computing Applications. 2023;37(6):683-705. doi:10.1177/10943420231201154

## List of abbreviations and definitions

| <b>Abbreviations</b> | <b>Definitions</b>  |
|----------------------|---|
| <b>ADM</b>           | Automated Decision-Making   |
| <b>AI</b>            | Artificial Intelligence   |
| <b>AI-RMF</b>        | Artificial Intelligence Risk Management Framework   |
| <b>AI-SIMD</b>       | AI-enabled Software In a Medical Device   |
| <b>ALTAI</b>         | Assessment List for Trustworthy Artificial Intelligence   |
| <b>ATMT</b>          | Algorithm-To-Model Transition'  |
| <b>AUC-ROC</b>       | Accuracy, Precision, Recall   |
| <b>CAs</b>           | Conversational Agents   |
| <b>CACAO</b>         | Collaborative Automated Course of Action Operations   |
| <b>CDBIO</b>         | Council of Europe's steering committee for human rights in the fields of biomedicine and health |
| <b>CIA</b>           | Confidentiality, Integrity and Availability   |
| <b>CIOMS</b>         | Council for International Organisations of Medical Sciences                                     |
| <b>CIS</b>           | Clinical Information Systems  |
| <b>COI</b>           | Conflicts Of Interest   |
| <b>CP</b>            | Clinical Pathway  |
| <b>CPG</b>           | Clinical Practice Guideline   |
| <b>CPP</b>           | Clinical Practice Protocol  |
| <b>CTM</b>           | Computational Theory of Mind  |
| <b>CRFM</b>          | Center for Research on Foundation Models  |
| <b>DPIA</b>          | Data Protection Impact Assessment   |
| <b>EC</b>            | European Commission   |
| <b>EGE</b>           | European Group on Ethics and New Technologies   |
| <b>EHR</b>           | Electronic Health Records   |
| <b>EIT</b>           | European Institute of Technology  |
| <b>ENISA</b>         | European Union Agency for Network and Information Security                                      |
| <b>EU</b>            | European Union  |
| <b>FDA</b>           | U.S. Food and Drug Administration   |
| <b>FOH</b>           | Future Of Health  |
| <b>FRAIA</b>         | Fundamental Rights and Algorithm Impact Assessment  |
| <b>FM</b>            | Foundation Models   |
| <b>FN</b>            | False Negatives   |
| <b>FP</b>            | False Positives   |
| <b>GDPR</b>          | General Data Protection Regulation  |
| <b>GHTF</b>          | Global harmonization task force   |
| <b>GMO</b>           | Genetically Modified Organisms  |
| <b>HCP</b>           | Health Care Professional  |
| <b>HHS</b>           | Health & Human Services   |
| <b>HIC</b>           | Human-In-Command  |
| <b>HITL</b>          | Human-In-The- Loop  |
| <b>HLEG</b>          | High-Level Expert Group   |

| <b>Abbreviations</b> | <b>Definitions</b>   |
|----------------------|--|
| <b>HOTL</b>          | Human-On-The-Loop  |
| <b>HTA</b>           | Health Technology Assessment   |
| <b>IA</b>            | Impact Assessment  |
| <b>IEEE</b>          | Institute of Electrical and Electronics Engineers  |
| <b>IMDRF</b>         | International Medical Device Regulators Forum  |
| <b>IT</b>            | Information Technology   |
| <b>IVD</b>           | In-vitro Diagnostic Medical Device   |
| <b>IVDR</b>          | In Vitro Medical Device Regulation   |
| <b>LLM</b>           | Large Language Models  |
| <b>MD</b>            | Medical Device   |
| <b>MDCE</b>          | Medical Device Clinical Evaluation Working Group   |
| <b>MDCG</b>          | Medical Device Coordination Group  |
| <b>MDR</b>           | Medical Device Regulation  |
| <b>MDSW</b>          | Medical Device Software  |
| <b>ML</b>            | Machine Learning   |
| <b>NGO</b>           | Non-Governmental Organisations   |
| <b>NIST</b>          | National Institute of Standards and Technology   |
| <b>OECD</b>          | Organisation for Economic Co-operation and Development   |
| <b>PANACEA</b>       | Protection and privAcy of hospital and health iNfrastructureS with smArt Cyber sEcURITY and cyber threat toolkit for dATA and people |
| <b>PATE</b>          | Private Aggregation of TEacher   |
| <b>PIA</b>           | Privacy Impact Assessments   |
| <b>RCT</b>           | Randomised Clinical Trials   |
| <b>RNTK</b>          | Right Not To Know  |
| <b>SAMD</b>          | Software As a Medical Device   |
| <b>SCM</b>           | Structured Causal Model  |
| <b>SOTA</b>          | State Of The Art   |
| <b>TRIPOD</b>        | Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis                                      |
| <b>TPR</b>           | True Positive Rate   |
| <b>UK</b>            | United Kingdom   |
| <b>UN</b>            | United Nations   |
| <b>USA</b>           | United States of America   |
| <b>VDE</b>           | Verband der Elektrotechnik Informationstechnik e.V.  |
| <b>WMA</b>           | World Medical Association  |
| <b>WHO</b>           | World Health Organisation  |
| <b>XAI</b>           | explainable AI   |

## **List of boxes**

|  |     |
|--|-----|
| <b>Box 1.</b> What in a nutshell is the ‘AI evidence pathway for health?’.....   | 20  |
| <b>Box 2.</b> Trust and trustworthiness are part of most documents on AI issued by public organisations.....   | 47  |
| <b>Box 3.</b> State of the art.....  | 57  |
| <b>Box 4.</b> Other risks not related to non-maleficence and AI safety:.....   | 71  |
| <b>Box 5.</b> Overlap of meaning between interpretability and explainability.....  | 125 |
| <b>Box 6.</b> Brief summary of EU legislations on defective products, developments concerning liability law and AI, lack of clarity around liability as an obstacle for AI adoption in healthcare..... | 148 |
| <b>Box 7.</b> Examples of open source toolkits for fairness, non-discrimination and equality.....  | 157 |
| <b>Box 8.</b> AI principles, ethical principles, ethics guidelines and recommendations issued by international organisations.....  | 194 |

## List of figures

|   |    |
|---|----|
| All figures and tables are our own production and have been created for this ontology.  |    |
| <b>Figure 1.</b> The five elements (see <b>Box 1</b> ) for enabling effective collaboration and evidence generation along the AI evidence pathway for health.....   | 11 |
| <b>Figure 2.</b> The community-bridging ontology as a fundamental element of the AI evidence pathway for health. The ontology has a two-layer design: ontology A lays out ten ethical principles and their translational concepts (three levels). These connect to ontology B (second layer) which describes fundamental socioethical, clinical and technical concepts, organised in 12 clusters.....   | 11 |
| <b>Figure 3. (a)</b> Number of translational concepts in ontology A for the 9 ethical principles plus 'trust' (150 concepts in total). <b>(b)</b> Number of concepts in ontology B per thematic clusters (157 concepts in total).....   | 16 |
| <b>Figure 4.</b> Schematic diagram of the contribution of the AI evidence pathway to AI management by providing a conceptual framework for collaboration of actors to generate evidence for trustworthy AI.....   | 21 |
| <b>Figure 5.</b> The AI evidence pathway as the nexus of people, trust, AI technologies and their adoption in health. People develop AI technologies and adopt them. People define conditions of trust and generate evidence supporting trustworthiness. People using AI generate real world evidence for safe innovation. ....   | 22 |
| <b>Figure 6.</b> Schematic depiction of the AI evidence pathway's five elements required for generating evidence on trust. Agents and communities are at the centre. They generate evidence on trustworthiness and safe innovation by collaborating along the pathway of life cycle stages and value chain elements. The evidence generation is supported by common concepts (ontologies). Evidence from real-world use should feed into new iterations of life cycle processes facilitating safe innovation. Left: AI governance and AI management as overarching concepts, drawing on the AI evidence pathway.....  | 24 |
| <b>Figure 7. (a)</b> The current interpretation gap between ethical principles formulating high-level conditions for trust and concepts to satisfy these conditions through evidence. <b>(b)</b> The ontology presented here contributes to closing this interpretation gap in the health area by providing nine granular ethical principles and 'translational concepts' that connect to fundamental concepts and aspects including the AI life cycle and the value chain. The ontology is intended to support community-bridging and collaboration to identify evidence needs and generate evidence on trustworthy AI. <b>(c)</b> The ontology can be interrogated either from Part A (9 ethical principles and translational concepts) or from part B (fundamental concepts). Both parts are connected via cross references..... | 29 |
| <b>Figure 8.</b> Schematic depiction of the two parts of this ontology: Part A builds on nine ethical principles which are unfolded in translational concepts. These point to fundamental concepts, structured in 12 thematic clusters of part B. Green dotted lines indicate cross connections. Fundamental concepts cover technical, clinical, ethical/philosophical, economic and legal/regulatory topics, covering the life cycle from conception and planning to assessment. Three ethical principles explicitly touch on the stage 'democratic debate & global aspects (e.g. on the future of AI use in healthcare, sustainability and solidarity issues) .....   | 30 |

**Figure 9.** (a) Derivation of the nine ethical principles of this ontology from the 11 consensus principles identified by Jobin et al., 2019. (b) Schematic depiction of the ethical principles and translational concepts for evidence generation to support trust and trustworthiness.....33

**Figure 10.** Four dimensions of trust as a prerequisite for adoption of AI technology in health. (a) Conditions for trustworthiness (=9 ethical principles), concerning both actors involved in building and deploying AI systems and the technology itself. (b) A priori trust in AI systems without sufficient oversight (e.g. → automation bias). A priori trust is comparably little debated. (c) Specific notions of trust in the context of health applications. This touches on a variety of issues, e.g. privacy of health data, patient-doctor relationship, informed consent. (d) situations of conflicts of interests, e.g. in case roles of developer and decision-maker concerning the acquisition and implementation of AI systems in a given healthcare setting coincide. ....45

**Figure 11.** Schematic depiction of the relationship of the concepts 'gains', 'added value', 'potential benefit' and 'real-world benefit' under the ethical principle of BENEFICENCE and along the AI evidence pathway.....57

**Figure 12.** Schematic depiction of the relationship of the concepts of NON-MALEFICENCE, → AI safety, → AI risk management, → AI management and → AI governance. N.B. The diagram is highly schematic. AI management, AI risk management and AI governance refer to **frameworks, processes or systems**. AI safety and controlled impacts to the **condition** of risks and impacts being adequately controlled. NON-MALEFICENCE to the **ethical obligation** to avoid harm, necessitating an **understanding of AI-associated risks and wider impact**. We understand NON-MALEFICENCE as an element of AI governance that feeds into AI management and in particular into AI risk management. ....69

**Figure 13.** Schematic Venn diagram of various communities in regard to transparency of information concerning AI in healthcare. The internal community (yellow; e.g. developers, businesses, organisations) will need to implement maximum transparency (e.g. to trace and understand potential failures). The community of compliance and HTA experts (blue) may need access to parts of confidential business information in order to assess agreement with legal and regulatory requirements and/or added value of an AI tool for health systems. The community of society in general, patients and users require transparent information on how to safely use an AI system and should be informed about problems and issues and how these were resolved.....109

**Figure 14.** Four elements required for intelligibility of AI systems and/or machine-learning models: (1) transparent information about the model, including how it, on a general level, transforms data into output or predictions; (2) interpretability of the prediction model affording a more detailed understanding of input-output causality or rules or generalisations used by the model; (3) scientific explanations providing a logical argument of transparency and interpretability elements and allowing predicting outcomes based on specific input data; (4) understandable communication of explanations to various target audiences.....123

**Figure 15.** Elements of intelligibility, part of the principle of transparency). Explicability is a composite concept that relates to intelligibility (and hence transparency) as well as accountability (and hence the preinciple of responsibility). ....130

**Figure 16.** Schematic depictions of types of reasoning that may be involved in explanatory arguments. (a) deductive reasoning (see deductive-nomological model), (b) inductive reasoning

|  |     |
|--|-----|
| <i>and (c) abductive reasoning, also referred to as inference to best possible explanation.</i>  |     |
| <i>Explainability approaches align often with abductive reasoning.....</i>   | 132 |
| <b>Figure 17. (a) Carnap's explication process, (b) Hempel &amp; Oppenheim's deductive-nomological model of an explanation; (c) logical explanation models for universal or statistical laws versus particular or general regularities (adapted from Salmon, 1989). .....</b>  | 134 |
| <b>Figure 18. (a) Scientific explanation according to the 'received view' proposed by Hempel &amp; Oppenheim, using a composite of antecedent conditions as well as laws / generalisation or well-confirmed hypotheses to deduce or infer an explanation. (b) Schematic depiction of hypothetico-deductive confirmation of hypotheses, which should not be confused with an explanation. (c) Hypothetico-deductive inference lens support to uncertain premises, based on a explanation found to be robust. Such inference is not an explanation. ....</b>   | 135 |
| <b>Figure 19. Mindmap of the term fairness and related terms.....</b>  | 153 |
| <b>Figure 20. Schematic depiction of actor communities and their simplified relation in regard to the life cycle and AI evidence pathway in health.....</b>  | 215 |
| <b>Figure 21. Schematic depiction of the ethical dimensions of bias and associated issues. Orange font: bias of people.....</b>  | 233 |
| <b>Figure 22. The AI evidence pathway is a conceptual framework for interconnecting all relevant elements for creating evidence for trustworthy AI. ....</b>   | 242 |
| <b>Figure 23. Proposed life cycle stages for AI in health. Stages 1-7 concern specific AI solutions. Stage 8 concerns broader discussions which however may radiate back to life cycle activities of individual products. A life cycle should be seen in the larger context of → AI governance, → AI management including → AI risk management activities. The → AI evidence pathway builds on this life cycle outline, connecting it to the → Value chain of AI and enabling community collaboration for evidence generation of trustworthy AI. ....</b>  | 245 |
| <b>Figure 24. Value chain dimensions. Value chain enablers in light green. Value chain assets in dark green.....</b>   | 247 |
| <b>Figure 25. Schematic depiction of the algorithm-to-model transition pathway (blue arrow) for → machine-learning based models/AI systems. Relationships to life cycle stages (yellow) and value chain elements (green) are indicated (*see Reina &amp; Griesinger, 2024b). ATMT should provide evidence on assumptions and decisions that are crucial for arriving at the final model. This includes model post-processing or optimisation steps based on validation and evaluation exercises. Other verification and validation steps (e.g. verification of design specifications of AI system, AI system validation, usability validation, clinical validation) are not part of ATMT. Terms in blue italic font correspond to ontology entries.....</b>  | 286 |
| <b>Figure 26. Highly schematic representation of key validation and evaluation activities feeding into clinical evaluation of AI systems used in healthcare and medicine. Dark grey stippled arrows: validation activities that may inform subsequent evaluation and validation. Light grey stippled arrows: information may contribute to clinical evaluation. Green arrows indicate evidence generated during dedicated prospective studies in the context of model evaluation and clinical validation. Blue arrows indicate main contributions from clinical validation and relevant post-market information, e.g. post-market surveillance, post-market clinical follow up. Yellow boxes = technical validation/testing. Green boxes: evaluation and validation generating clinically relevant</b> |     |

evidence. Orange boxes: information from the post-market space. Note that various impact assessments that may be required are not shown ..... 323

**Figure 27.** Highly schematic representation of key validation and assessment activities that, depending on AI system, may need to be considered for the continuous evaluation of AI systems that are not used in healthcare and medicine but for other applications such as health research, health system management and public health. Dark grey stippled arrows: validation and testing activities that may inform subsequent AI system validation and usability validation. Light grey stippled arrows: information may contribute to continuous evaluation of the AI system. Blue arrows indicate contributions of validation exercises and post-market information to the continuous evaluation of the AI system. Yellow boxes = technical validation/testing. Purple box: information from the post-market space. Note that various impact assessments that may be required are not shown..... 325

**Figure 28.** Schematic depiction of the term 'use' and main related concepts. There are four clusters: 1) user, 2) application in health (as outlined in WHO, 2021). 3) Information for use, 4) Use context, use environment, real-world use. ..... 340

## List of tables

|  |     |
|--|-----|
| <b>Table 1.</b> When developing the ontology we tried to address possible omissions of topics in ethical guidelines as suggested by Hagendorff (2020) as well as additional aspects. The table summarizes our considerations. Ethical principles in red font, translational concepts in black font.  | .34 |
| <b>Table 2.</b> Relationship between the seven key requirements of the “ethics guidelines for trustworthy AI” by the EU Commission’s independent high-level expert group (EC HLEG, 2019) and concepts outlined in this ontology. The definitions of key requirements are from Recital 27 of the EU’s AI Act, except for ‘accountability’. Ethical principles (part A) in red font and thematic clusters (part B) in green font. Concepts in black. | .36 |

## Annexes

### Annex 1 – Tools for clinical studies and evaluation of AI in healthcare

#### EQUATOR network: enhancing the quality and transparency of health research

| Concept description   |
|---|
| <p>The network “Enhancing the quality and transparency of health research (EQUATOR)” was established in 2008. It aims to enhance the value, reliability and transparency of published research on health (from lab bench to clinical studies) by promoting and providing relevant guidelines (both published and under development) for the reporting a variety of health-related studies (e.g. preclinical animal studies, systematic reviews, clinical studies) as well as other tools, e.g. on scientific writing.</p> <p>Several clinical guidelines as well as their AI-related extensions (e.g. STARD, DECIDE, TRIPOD) described briefly in this Annex are listed on and accessible from the EQUATOR’s online database.</p> |
| Explanatory note and references   |
| <p>EQUATOR is a useful one-stop shop for identifying and downloading relevant expert-based guidelines that provide practical guidance and are free and transparent to all actors.</p> <ul style="list-style-type: none"><li>• Website of EQUATOR network. Online: <a href="https://www.equator-network.org/">https://www.equator-network.org/</a> (accessed 2024.08.06)</li></ul>   |

#### STARD-AI

| Concept description   |
|---|
| <p>STARD-AI is an extension of STARD (<i>‘standards for reporting of diagnostic accuracy studies’</i>) used to <b>report diagnostic accuracy studies</b>, either at development stage or as an offline validation in clinical settings.</p> <p>STARD-AI focuses on diagnostic accuracy and aims to aid the comprehensive reporting of research that use AI techniques to assess diagnostic test accuracy and performance.</p> <p>This can account for either single or combined test data, which often consists of either (1) imaging data (e.g., CT scans), (2) pathological data (e.g., digitised specimen slide) or (3) reporting data (e.g., electronic health records).</p>  |
| Explanatory note  |
| <p>The STARD 2015 statement is a widely accepted set of reporting standards for diagnostic accuracy studies. STARD was developed to improve the completeness and transparency of studies investigating diagnostic accuracy. Notably, STARD was not designed to address various methodological issues and inconsistencies, including differing methodological interpretation (e.g., the use of external validation datasets (→ <b>model validation</b>), complexities of datasets and comparison to human performance), the lack of standardized nomenclature and commo, consistent understanding, as well as the heterogeneity of outcome measures (e.g., area under the receiver operating characteristics (AUROC; → <b>receiver operating characteristic</b>), → <b>sensitivity</b>, positive predictive value or F1 score).</p> <p>In order to tackle these problems, the STARD-AI Steering Group has prepared an AI-specific extension to the STARD statement (STARD-AI) that aims to focus upon the specific reporting of AI diagnostic accuracy studies. This work aims to complement other methodological extensions for AI tools, developed by the EQUATOR (Enhancing Quality and Transparency of Health Research) network program, such as →</p> |

**CONSORT-AI** (Consolidated Standards of Reporting Trials), → **SPIRIT-AI** (Standard Protocol Items: Recommendations for Interventional Trials) and → **TRIPOD-ML** (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis).

References:

- Sounderajah et al (2020, 2021)
- Bossuyt et al (2015)

## TRIPOD-AI

### Concept description

TRIPOD-AI is an extension of TRIPOD ('transparent reporting of a multivariable prediction model for individual prognosis or diagnosis') **used to report prediction models** (diagnostic or prognostic) development, validation and updates.

### Explanatory note

Models that predict clinical outcomes are broadly categorised as those that estimate the probability of the presence of a particular outcome (diagnostic) or whether a particular outcome (e.g. event) will occur in the future (prognostic).

The TRIPOD statement was published in 2015, proposing minimum reporting recommendations for studies developing or evaluating the performance of a prediction model. Traditionally, prediction models have been developed using regression-based methods, typically logistic regression for short-term outcomes and Cox regression for longer-term outcomes. Numerous reviews have observed that studies describing the development and validation (including updating) of a prediction model often fail to report key information to help readers judge the methods and have a complete, transparent and clear picture of the model's predictive accuracy and other relevant details such as the target population and the content of the model itself. In response to this, in 2015, the TRIPOD Statement was published.

Machine learning methods are increasingly developed and used for specific real-world diagnostic or prognostic problems. TRIPOD-AI aims to promote the complete, accurate, and transparent reporting of studies that develop a → **machine learning**-based prediction model or evaluate its performance. Complete reporting will facilitate study appraisal, → **model evaluation**, and model implementation (see also → **actionability**).

References:

- Collins et al. (2015, 2021, 2024)

## PROBAST-AI

### Concept description

PROBAST-AI is an extension of the PROBAST tool for prediction model studies (PROBAST = prediction model risk of bias assessment tool). PROBAST-AI and → **TRIPOD-AI** are intended to **improve reporting and clinical appraisal of prediction model studies** for diagnosis and prognosis using → **machine learning** ("artificial intelligence") techniques. PROBAST and PROBAST-AI have been developed to assess the risk of bias or distortion as well as applicability of prediction models, compared to their development setting (→ **applicability and limitations**).

### Explanatory note

PROBAST-AI is currently being refined.

References:

- Collins GS et al (2021)

## DECIDE-AI

### Concept description

The '*developmental and exploratory clinical investigations of decision support systems driven by Artificial Intelligence*' (DECIDE-AI) is an extension of the established DECIDE tool. DECIDE-AI aims at the **early evaluation of AI systems**, with an emphasis on the human users (→ **usability**). DECIDE-AI focuses on four dimensions: 1) performance of AI tools when first used with humans in actual clinical settings (typically small scale), 2) assessment of the safety profile of the AI tool before routine deployment, 3) evaluation of "human factors" of the AI tool (so-called 'ergonomics'), 4) preparatory steps towards large-scale → **clinical investigations or trials**.

See also entry on → **model evaluation**.

### Explanatory note

Comprehensive early evaluation, with an emphasis on the human users, is a necessary bridge from the development of clinical decision support systems driven by AI to their clinical implementation. Early-stage studies are also important steppingstones toward the large-scale comparative trials needed to generate robust evidence about these systems' effectiveness.

Early evaluation studies in small-scale clinical settings may bridge the gap between

- the *in silico* algorithm development/validation (addressed by the → **TRIPOD-AI** and → **STARD-AI** statements) and
- large-scale clinical trials evaluating AI interventions (addressed by the → **CONSORT-AI** statement).

Conceptually, this early evaluation step can be compared to a phase 1/2 trial for drug development or (a much closer analogy, given the relationship between users' characteristics and the intervention's effectiveness) IDEAL stage 2a/2b for surgical innovation.

### References:

- The DECIDE-AI Steering Group (2025)
- Vasey et al (2022)

## SPIRIT-AI

### Concept description

SPIRIT-AI is the acronym for '*standard protocol items: recommendations for interventional trials-artificial intelligence*'. SPIRIT-AI is a proposed **reporting guideline for clinical trial protocols evaluating interventions with an AI component**.

SPIRIT-AI is an extension of the SPIRIT 2013 statement ('*the standard protocol items: recommendations for interventional trials*') that aims to improve the completeness of clinical trial protocol reporting, by providing evidence-based recommendations for the minimum set of items to be addressed.

SPIRIT-AI was developed in parallel with its companion statement for trial reports: → **CONSORT-AI** (Consolidated Standards of Reporting Trials-Artificial Intelligence).

### Explanatory note

There is a growing recognition that interventions involving artificial intelligence tools need to undergo rigorous, prospective evaluation to demonstrate their impact on → **health outcomes**.

SPIRIT-AI intends to promote → transparency and completeness for clinical trial protocols for AI interventions. To this end, SPIRIT-AI is particularly aimed at investigators planning or conducting clinical trials. It may also serve as guidance for developers of AI interventions at earlier → validation stages of an AI system (→ model validation, → model evaluation).

Further, it is hoped that SPIRIT-AI will assist peer reviewers and other relevant experts to understand, interpret, and critically appraise the design and risk of → bias for a planned clinical trial.

#### References:

- Rivera et al (2020)
- Chan et al (2013)

## CONSORT-AI

### Concept description

CONSORT-AI ('consolidated standards of reporting trials–artificial intelligence') is a **reporting guideline for clinical trials evaluating interventions with an AI component**. CONSORT-AI focuses on effectiveness and safety, and it is used to report randomised controlled trials (RCT) evaluating AI systems as interventions (large scale, summative evaluation), independently of the → healthcare modality of the AI system (e.g. diagnostic, prognostic, therapeutic).

CONSORT-AI is an extension of the 2010 update of the initial CONSORT statement (published in 1996) which provided recommendations for the publication of randomised controlled clinical trials, the gold standard to assess healthcare intervention.

CONSORT-AI was developed in parallel with its companion statement for clinical trial protocols: SPIRIT-AI (standard protocol items: recommendations for interventional trials–Artificial Intelligence).

### Explanatory note

There is a growing recognition that interventions involving artificial intelligence need to undergo rigorous, prospective evaluation to demonstrate their impact on → health outcomes. CONSORT-AI recommends that investigators provide clear descriptions of the AI intervention, including instructions and skills required for use, the setting in which the AI intervention is integrated, the handling of inputs and outputs of the AI intervention, the human–AI interaction and provision of an analysis of error cases. CONSORT-AI is aimed at promoting transparency and completeness in reporting clinical trials for AI interventions.

#### References:

- Liu et al. (2020)
- Schulz et al (2010)
- Begg et al (1996)

## CLAIM

### Concept description

The CLAIM ('checklist for Artificial Intelligence in medical imaging') is a **guideline aiming to promote the complete and consistent reporting of AI science in medical imaging**.

CLAIM has been initially developed in 2020 after the STARD guideline and has been extended to address applications of AI in medical imaging that include classification, image reconstruction, text analysis, and workflow optimisation.

In its initial version (2020) CLAIM was based on a checklist consisted of 42 items which serves a "best practice" to guide authors in presenting their research in the scientific field of AI in medical imaging.

In order to address AI's rapid scientific evolution and promote the consistent reporting of scientific advances of AI in medical imaging an updated version of CLAIM was presented in 2024 with an updated checklist comprising 44 items (checkpoints) where each CLAIM item now has three options, i.e. Yes, No, and Not Applicable (NA).

The 2024 CLAIM update provides a best practice checklist to promote transparency and reproducibility of medical imaging AI research.

#### Explanatory note

CLAIM has been adopted widely by several medical specialties that involve imaging and AI.

In its 2024 updated version the CLAIM guideline discourages authors from using the term “→ validation” because of its ambiguity. Instead, the CLAIM 2024 guideline encourages authors to use the terms “internal testing” and “external testing” to describe the process of testing against a held-out subset of training data and a completely external dataset (such as one from another site), respectively. Furthermore, the 2024 CLAIM update advises authors to use the term “validation” with caution and to consider using “tuning” or “model optimisation.” (see → model validation).

Another key feature of the CLAIM 2024 update is that authors are encouraged to use the term “reference standard” instead of the “ground truth” in order to indicate the benchmark against which an AI model’s performance is measured. This way connotations and ambiguities of jargon-like terms such as “ground truth” and “gold standard,” are avoided while there is alignment with the language of other scientific reporting guidelines e.g. the STARD.

Finally, in order to assist authors and reviewers of AI manuscripts in medical imaging during their reporting and review process, the 2024 CLAIM checklist encourages authors to record the page numbers of their manuscript where relevant information can be found concerning the checklist items (in cases of answering YES to the CLAIM item). Similarly, for those checklist items marked as No or NA, authors are encouraged to provide an explanation.

#### References:

- Tejani AS, et al (2024b).
- Mongan J, et al. (2020)

## MI-CLAIM

#### Concept description

The MI-CLAIM ('minimum information about clinical Artificial Intelligence modelling') is a **general reporting guideline for medical artificial intelligence studies**. The MI-CLAIM guideline targets mainly **medical-algorithm designers, repository managers**, manuscript writers and readers, journal editors, and model users.

MI-CLAIM has been developed to enable the assessment of clinical impact (including fairness and bias) and the rapid replication of the technical design process of legitimate clinical AI studies.

The MI-CLAIM reporting guideline comprises 17 binary (yes/no) items organized into 6 domains: study design, data optimisation, model optimisation, model performance, model examination, and reproducibility.

#### Explanatory note

The goal of the MI-CLAIM guideline is to establish a documentation standard that can serve clinical scientists, data scientists, and the clinicians of the future who will be using AI tools in the clinical setting. Thus, in order to assist the reporting and review process, the MI-CLAIM guideline invites authors to record on the MI-CLAIM checklist the page numbers of their manuscript where relevant information can be found regarding the checklist items.

#### References:

- Norgeot, B. et al. (2020)

## FUTURE-AI

### Concept description

The FUTURE-AI guideline aims at the **development and deployment of trustworthy and ethical AI in healthcare**. The guideline is a proposed framework of about 30 “best practices” that address technical, clinical, socioethical and legal dimensions. The 30 practices are rooted in six “guiding principles” (fairness, universality, traceability, usability, robustness, and explainability; these make up the acronym ‘future’). The principles represent a mix of ethical principles (*fairness* – covering health equity, inclusiveness, non-discrimination and absence of bias and *universality* – the latter covering generalisability, adaptability and interoperability) as well as technical requirements that are, in the guideline, not further related to ethical principles (*traceability, usability, robustness, explainability*). The guideline is the work of a consortium of multidisciplinary experts from various continents and a variety of disciplines, e.g. data science, medical research, clinical medicine, computer engineering, medical ethics, social sciences as well as certain data domains, e.g. radiology, genomics, mobile health, electronic health records, surgery, pathology. Work on the guideline started in 2021. The guideline was published in early 2025.

### Explanatory note

Several guidelines have been developed so far focusing on providing recommendations for the reporting of AI studies regarding different medical domains or clinical tasks e.g. TRIPOD-AI, CONSORT-AI, DECIDE-AI, PROBAST-AI. However, according to FUTURE-AI, these guidelines do not provide best practices for the actual development and deployment of the AI tools used in the relevant studies.

The FUTURE-AI framework is based on a risk-based approach and proposes, as previously [WHO \(2021a\)](#) and the EU high-level expert group ([EU HLEG, 2019](#)), the assessment of AI application specific risks and challenges early in the process (e.g. risk of discrimination, lack of generalisability, data drifts over time, lack of acceptance by end users, potential harm for patients, lack of transparency, data security vulnerabilities, ethical risks). As other guidelines before, FUTURE-AI recommends the continuous engagement with multidisciplinary stakeholders to understand specific needs, risks, and solutions of each AI application as this is considered crucial for the thorough exploration of all possible risks and factors that might lead to reduced trust in a specific AI tool. FUTURE-AI seeks to be a living framework, thus a dedicated webpage ([www.future-ai.eu](http://www.future-ai.eu)) has been created in order to collect expert opinions that might lead to future revisions of the FUTURE-AI guideline.

### References:

- Future-AI consortium (2023, 2025)

## QUADAS-AI

### Concept description

QUADAS ('*quality assessment of diagnostic accuracy studies*') and later QUADAS-2 were developed as tools to aid the evaluation of methodological quality within primary diagnostic accuracy studies, aiming particularly at the assessment of → **bias** and applicability. The use of QADAS-2 is encouraged by PRISMA 2020 guidance ([Page et al., 2021](#)). However, QUADAS-2 has not been developed with a view to the specific challenges related to the quality assessment of AI diagnostic accuracy studies, e.g. terminology, → **bias** within the class of study or → **algorithmic bias**. In order to tackle these specifics of AI diagnostic accuracy studies, the QUADAS-AI tool has been proposed, which is intended as an extension of QUADAS-2 and QUADAS-C for comparative diagnostic accuracy studies ([Yang et al., 2021](#)).

### Explanatory note

The future development of QUADAS-AI will be conducted in three stages. Stage 1 covers the organisational steps e.g. the establishment of the project team and the steering team while stage 2 will focus on

the item generation process based on a modified Delphi consensus methodology following a) a mapping review, b) a meta-research study, 3) a scoping survey of international experts, and d) a patient and public involvement and engagement exercise. In the final third stage specific dissemination strategies of the finalised QUADAS-AI tool, will be implemented toward academic, policy, regulatory, industry, and public stakeholders.

References:

- Guni et al., (2024)
- Sounderajah et al., (2021b)

## CORE-MD clinical risk score

### Concept description

Current guidance on clinical evaluation of medical device software under the EU's medical devices Regulation ([EU, 2017a](#)) and performance evaluation under the in vitro diagnostics medical devices Regulation ([EU, 2017b](#)) does not fully address specific clinical evidence needs in relation to → AI-enabled medical device software ([EU Medical Devices Coordination Group, 2020a](#)).

The CORE-MD research consortium ('coordinating research and evidence for medical devices', a EU-funded horizon 2020 project) established a task force with the objective to recommend possible methodological principles for the clinical and performance evaluation of AI-enabled medical device software, taking into account key principles under the EU's MDR Regulation (e.g. benefit-risk ratio), the overall regulatory requirements of the Regulations and the clinical evidence needed for the trusted use of AI-based medical devices. A proposal for a clinical risk score allowing the stratification of clinical/performance evidence needs was published in 2025. According to CORE-MD, the application of this set of regulatory recommendations in the clinical evaluation of AI-based MDSW could balance the beneficial impacts of AI use in healthcare against the potential negative effects from inappropriate use or misuse.

### Explanatory note

The CORE-MD risk-benefit approach is based on the development of a risk-based scoring system for medical devices incorporating artificial intelligence elements e.g. machine learning techniques or algorithms. The total clinical risk score corresponds to the sum of sub-scores in three evaluation domains: (1) clinical performance score (see also → [clinical performance](#)), (2) valid technical performance score (→ [validation](#)) and (3) valid clinical association score of the AI-based MDSW under evaluation (→ [valid clinical association / scientific validity](#)). The total score can be used as a decision-making support tool for considering the level clinical risk – and, correspondingly, → [clinical evidence](#) required to bolster trust in the tool. For example, high-risk scores may alert to the possible need for conduct of extensive clinical investigations prior to regulatory approval, while lower scores might identify low-risk AI-based medical devices which may require less extended pre-market clinical evaluation, while relying stronger on post-market activities and evidence. The CORE-MD proposal, conceptually, reflects a central tenet of the EU's medical device Regulation: the acceptability of risks of devices is defined on the basis of a benefit-risk ratio, generated during → [clinical evaluation](#) (e.g. EU's medical device Regulation: Annex I, Chapter 1, point 2; EU, 2017a); see also entry on → [AI risk management](#).

References:

- Rademakers et al. (2025)

## Annex 2 - Detailed table of contents

|   |           |
|---|-----------|
| <b>Abstract .....</b>   | <b>2</b>  |
| <b>How to navigate this document.....</b>   | <b>3</b>  |
| <b>Foreword.....</b>  | <b>4</b>  |
| <b>Executive summary .....</b>  | <b>5</b>  |
| <b>Acknowledgements .....</b>   | <b>9</b>  |
| Authors .....   | 9         |
| Author contributions .....  | 9         |
| Use of artificial intelligence .....  | 9         |
| <b>1     Introduction .....</b>   | <b>10</b> |
| 1.1     About this ontology .....   | 10        |
| 1.1.1     The ontology as a foundation of an ‘AI evidence pathway for health’.....              | 10        |
| 1.1.2     Application areas of AI in health.....  | 12        |
| 1.1.3     About the term ‘ontology’ .....   | 12        |
| 1.1.4     Objective of this ontology .....  | 12        |
| 1.1.5     In scope and out of scope aspects of this ontology .....                              | 13        |
| 1.1.6     Structure of this ontology: two interlinked parts .....                               | 14        |
| 1.1.7     Term layout, formatting and references .....  | 17        |
| 1.1.8     Practical applications of the ontology .....  | 18        |
| 1.1.9     How to use this ontology .....  | 19        |
| 1.2     The ‘AI evidence pathway’: collaboration for evidence .....                             | 20        |
| 1.2.1     The AI evidence pathway: overview .....   | 20        |
| 1.2.2     AI evidence pathway: a people-centric and collaborative framework.....                | 22        |
| 1.2.3     The five dimensions of the evidence pathway.....                                      | 23        |
| 1.3     Our motivation for the AI evidence pathway .....  | 25        |
| 1.3.1     No international consensus on ethical or AI principles .....                          | 25        |
| 1.3.2     Gap between ethical principles and practical concepts.....                            | 26        |
| 1.3.3     Most guidelines and ethics documents are not tailored to health.....                  | 27        |
| 1.3.4     Addressing the three challenges through the AI evidence pathway for health .....      | 27        |
| <b>2     Methodological approach for developing this ontology .....</b>                         | <b>31</b> |
| 2.1     Premises .....  | 31        |
| 2.2     Derivation of nine ethical principles plus trust as a desideratum .....                 | 32        |
| 2.3     Fundamental concepts associated with ethical principles and translational concepts..... | 40        |
| <b>3     Ontology A: Ethical principles and translational concepts.....</b>                     | <b>42</b> |
| A.0     Trust and trustworthiness .....   | 42        |
| Trust and trustworthiness: translational concepts .....   | 49        |
| Conditions for trust and trustworthiness: actors and AI systems .....                           | 50        |
| A priori trust in AI systems.....   | 51        |
| Specific aspects of trust in the health domain .....  | 51        |
| Trust issues related to situations of conflict of interest (Col) .....                          | 51        |
| A.1     Beneficence .....   | 53        |
| Beneficence: translational concepts .....   | 56        |
| Generating and evaluation evidence on benefits .....  | 56        |
| Added value .....   | 57        |
| Real-world benefits.....  | 58        |
| Potential benefits.....   | 59        |
| Gains .....   | 59        |
| General output quality gains .....  | 59        |
| Precision gain .....  | 59        |
| Accuracy gain .....   | 59        |
| Consistency gain.....   | 59        |
| Non-bias gain & equitable bias gains .....  | 59        |
| General gains of operational efficiency and usability .....                                     | 60        |

|  |    |
|--|----|
| Time-to-delivery gain.....   | 60 |
| Reduced burden for HCP .....   | 60 |
| Usability gain.....  | 60 |
| Economic gains .....   | 60 |
| Economic gain - person time.....   | 60 |
| Economic gain: investments.....  | 60 |
| Economic gain: returns .....   | 60 |
| Economic gain: less need for non-human resources.....  | 60 |
| Clinical gains.....  | 61 |
| Clinical benefits for patient.....   | 61 |
| Reducing heuristic decision-making.....  | 61 |
| Well-being gains .....   | 61 |
| Patient well-being.....  | 61 |
| Higher degree of autonomy / independence of patients with chronic diseases and/or debilitating conditions..... | 61 |
| Healthcare professional well-being.....  | 61 |
| Improved patient-physician relationship.....   | 61 |
| Healthcare gains .....   | 62 |
| General healthcare management gains.....   | 62 |
| Healthcare adaptation gains .....  | 62 |
| Healthcare accessibility gains.....  | 62 |
| Healthcare equality gains.....   | 62 |
| Healthcare equity gains .....  | 62 |
| Global healthcare equality & equity gain(s) .....  | 62 |
| Healthcare precision gain .....  | 62 |
| Healthcare quality gain.....   | 62 |
| Healthcare efficiency gain .....   | 62 |
| Healthcare gain - unmet medical or patient needs .....   | 63 |
| Healthcare gain - personalisation of prevention/treatment.....   | 63 |
| Gain in regard to complex ethical considerations in healthcare.....  | 63 |
| Social and societal gains, including public health.....  | 63 |
| Inclusivity gain.....  | 63 |
| Societal gain.....   | 63 |
| Gains for population / public health.....  | 63 |
| General health system management and planning gain .....   | 64 |
| Socioeconomic gain.....  | 64 |
| Transparency gain.....   | 64 |
| Digital infrastructure gains .....   | 64 |
| Data access or collection, wrangling.....  | 64 |
| Synthetic data.....  | 64 |
| Decommissioning.....   | 65 |
| Health research gains .....  | 65 |
| Environmental gain .....   | 65 |
| A.2 <i>Non-maleficence</i> .....   | 66 |
| Non-maleficence: translational concepts .....  | 71 |
| Generating and evaluating evidence on risks .....  | 71 |
| Risks related to dignity, freedom and autonomy.....  | 71 |
| Risks related to data privacy of personal information .....  | 72 |
| Risks related to transparency .....  | 73 |
| Risks related to responsibility.....   | 73 |
| Risks related to bias .....  | 74 |
| Risks related to insufficient robustness / resilience and generalisability.....                                | 75 |
| Robustness / resilience: use context and use environment variations & drift issues.....                        | 75 |
| Generalisability: shortcut learning.....   | 76 |
| Cybersecurity: risks for value chain assets and knock-on effects on safety .....                               | 77 |
| Risks related to value chain elements .....  | 78 |
| Risks relating to technical integration and interoperability issues.....                                       | 80 |
| Risks related to integration into workflows .....  | 81 |
| Risks related to pre-deployment evidence gaps.....   | 82 |

|   |            |
|---|------------|
| Risks related to insufficient post-deployment monitoring / post-market surveillance .....             | 82         |
| Monitoring incidents / adverse events.....  | 82         |
| Monitoring performance, effectiveness, efficiency and bias .....                                      | 83         |
| Monitoring drifts / shifts.....   | 83         |
| Monitoring usability and usability-related errors .....   | 83         |
| Generating evidence on wider impacts .....  | 83         |
| Impact assessments.....   | 83         |
| <b>A.3     <i>Dignity, freedom and autonomy</i></b> .....   | <b>84</b>  |
| Dignity: translational concepts .....   | 87         |
| Respecting patient primacy .....  | 87         |
| AI and the development of healthcare.....   | 87         |
| Open, democratic and evidence-based debate .....  | 88         |
| Upholding a trustful patient-physician relationship.....  | 89         |
| Avoiding automation bias .....  | 90         |
| Avoiding automation complacency .....   | 91         |
| Deskilling.....   | 91         |
| Ensuring the means for free and informed consent.....   | 92         |
| Right to know and right not to know.....  | 94         |
| Right to know if AI system employed.....  | 94         |
| Medical privacy / health privacy.....   | 95         |
| Consent concerning collection of personal data / medical information.....                             | 95         |
| Right of intelligible information about personal data being processed.....                            | 96         |
| Right to demand rectification or erasure of data processed not in line with relevant provisions ..... | 97         |
| <b>A.4     <i>Privacy protection</i></b> .....  | <b>98</b>  |
| Privacy protection: translational concepts .....  | 101        |
| Data protection.....  | 101        |
| Lawful, legitimate and fair use of data .....   | 101        |
| Impact assessments: risk identification, data protection, benefits.....                               | 102        |
| Management of right of access (e.g. EU's GDPR, Data Act) .....  | 102        |
| Data governance, data management & data accountability.....   | 104        |
| Data security.....  | 105        |
| CIA principle from a privacy protection perspective .....   | 105        |
| Prevention of unauthorised access (e.g. servers, devices, platforms, cloud).....                      | 106        |
| Privacy-preserving techniques for AI development .....  | 107        |
| <b>A.5     <i>Transparency</i></b> .....  | <b>108</b> |
| Transparency: translational concepts.....   | 112        |
| Transparency of organisations and actors providing/deploying AI systems .....                         | 112        |
| Communication.....  | 112        |
| Disclosure .....  | 113        |
| Transparency of human-AI interaction.....   | 113        |
| Traceability.....   | 114        |
| Failure transparency .....  | 114        |
| Transparency of AI systems.....   | 115        |
| Evidence on algorithm-to-model transition.....  | 116        |
| Evidence on data .....  | 116        |
| Evidence about frameworks and processes used.....   | 116        |
| Evidence on intended use, applicability and limitations .....   | 117        |
| Evidence on interoperability and value chain elements.....  | 118        |
| Information on training requirements .....  | 119        |
| Intelligibility .....   | 119        |
| Interpretability and explainability .....   | 124        |
| Explicability .....   | 129        |
| Explanations of AI systems and their outcomes .....   | 131        |
| Understandable explanations .....   | 137        |
| <b>A.6     <i>Responsibility</i></b> .....  | <b>138</b> |
| Responsibility: translational concepts .....  | 141        |
| Legal and regulatory compliance.....  | 141        |
| Acting with integrity.....  | 141        |

|   |            |
|---|------------|
| Ethics code & governance / management frameworks .....                                    | 141        |
| Quality culture and risk management.....  | 142        |
| Correcting problems and failures, including necessary communication .....                 | 142        |
| Peer review and community discourse.....  | 142        |
| Accountability .....  | 143        |
| Accountability structures, attribution of (distributed) responsibilities .....            | 143        |
| Auditability and auditing.....  | 144        |
| Ensuring human agency and oversight.....  | 144        |
| Responsiveness .....  | 146        |
| Contestability and challenge .....  | 146        |
| Remedy and redress.....   | 147        |
| Liability.....  | 147        |
| A.7 <i>Fairness</i> .....   | 151        |
| Fairness: translational concepts .....  | 156        |
| Avoiding discrimination and discriminatory bias.....                                      | 156        |
| Looking out for and avoiding discriminatory bias .....                                    | 157        |
| Monitoring and mitigation of possible discrimination throughout the evidence pathway..... | 158        |
| Health equality & health equity.....  | 159        |
| Unavoidable trade-offs .....  | 160        |
| Universal versus targeted design .....  | 161        |
| A.8 <i>Solidarity</i> .....   | 162        |
| Solidarity: translational concepts.....   | 165        |
| Open, democratic and solidary society, culture of dialogue .....                          | 165        |
| Taking vulnerable groups into account “by design” .....                                   | 165        |
| Considering patients with rare diseases / conditions.....                                 | 166        |
| Considering mental health impacts of AI exposure.....                                     | 166        |
| Accessibility of data and infrastructure for AI development.....                          | 166        |
| Avoiding colonial structures; reducing the global north-south divide.....                 | 166        |
| Job impacts, skills, training.....  | 167        |
| Job impacts .....   | 167        |
| Training and skills .....   | 167        |
| Safe innovation: community bridging, collaboration, education .....                       | 168        |
| Private-public partnerships .....   | 169        |
| A.9 <i>Sustainability</i> .....   | 171        |
| Sustainability: translational concepts .....  | 173        |
| Environmental sustainability of AI system throughout the life cycle .....                 | 173        |
| Sustainability of health systems using AI .....   | 173        |
| <b>4     Ontology B: Fundamental concepts .....</b>                                       | <b>174</b> |
| <b>B.1     <i>AI and AI systems</i>.....</b>  | <b>174</b> |
| Artificial intelligence (AI) .....  | 174        |
| AI as a scientific field.....   | 175        |
| Computational theory of mind (CTM).....   | 176        |
| AI systems .....  | 178        |
| AI system component / part .....  | 179        |
| AI-enabled medical device software .....  | 180        |
| AI-enabled software in a medical device (AI-SIMD) .....                                   | 182        |
| AI-enabled software as a medical device (AI-SAMD) .....                                   | 182        |
| AI system output .....  | 182        |
| AI typology .....   | 183        |
| AI technique .....  | 184        |
| Conversational agents .....   | 185        |
| <b>B.2     <i>Ethics, AI ethics, governance, management</i> .....</b>                     | <b>187</b> |
| Ethical principles.....   | 187        |
| Bioethics .....   | 189        |
| Principilism .....  | 190        |
| AI ethics .....   | 191        |
| AI principles and AI ethics guidelines .....  | 192        |
| Ethical evaluation of AI.....   | 195        |

|  |     |
|--|-----|
| AI impact assessment (AI-IA) .....                                       | 197 |
| Fundamental rights and algorithm impact assessments.....                 | 198 |
| Alignment.....   | 199 |
| Ethics code .....  | 200 |
| AI governance.....   | 201 |
| AI management .....  | 204 |
| AI risk management.....  | 205 |
| Risk.....  | 207 |
| AI safety .....  | 208 |
| Human-centric AI .....   | 211 |
| Research involving human subjects: ethical principles & guidelines ..... | 212 |
| <i>B.3 Al actors and communities.</i> .....                              | 214 |
| AI actors.....   | 214 |
| AI practitioners .....   | 215 |
| Actors as defined in the EU's AI Act.....                                | 217 |
| Users of AI in the health domain.....                                    | 217 |
| AI value chain actors .....  | 219 |
| Compliance experts .....   | 220 |
| Health technology assessment (HTA) experts .....                         | 220 |
| Health policy makers, health economists, health system managers.....     | 221 |
| <i>B.4 Agency, autonomy and automation.</i> .....                        | 222 |
| Agency .....   | 222 |
| Human agency .....   | 223 |
| Human oversight.....   | 224 |
| Affective computing.....   | 224 |
| Autonomy.....  | 225 |
| Human primacy.....   | 226 |
| Corrigibility.....   | 227 |
| Patient primacy .....  | 228 |
| Machine agency .....   | 228 |
| Automation .....   | 229 |
| Augmentation.....  | 229 |
| <i>B.5 Bias, heuristics, drift &amp; shift.</i> .....                    | 231 |
| Bias .....   | 231 |
| Heuristics.....  | 234 |
| Shortcut learning.....   | 236 |
| Algorithmic bias .....   | 237 |
| Intrinsic incompatibilities or 'trade-offs' .....                        | 237 |
| Drift / shift in machine learning .....                                  | 238 |
| Concept drift / shift.....   | 239 |
| Data drift / shift.....  | 239 |
| Distributional drift / shift .....                                       | 240 |
| <i>B.6 AI Evidence pathway, AI life cycle and AI value chain.</i> .....  | 241 |
| AI Evidence pathway for health.....                                      | 241 |
| Life cycle of AI in health.....  | 243 |
| Value chain of AI.....   | 246 |
| Health technology assessment.....  | 248 |
| Post-deployment monitoring.....  | 249 |
| Decommissioning / Retirement.....  | 250 |
| <i>B.7 Data.</i> .....   | 252 |
| Dataset.....   | 252 |
| Development data.....  | 252 |
| Data provenance of development data.....                                 | 252 |
| Training data.....   | 253 |
| Validation data .....  | 253 |
| Testing data .....   | 254 |
| Input data .....   | 254 |
| Output and output data .....   | 255 |
| Data quality.....  | 255 |

|  |            |
|--|------------|
| Data quality metrics.....  | 257        |
| Post-deployment input data .....                                     | 259        |
| Data modality .....  | 259        |
| Nature of data.....  | 260        |
| Datasheets for datasets.....   | 260        |
| Synthetic health data.....   | 260        |
| Data privacy .....   | 261        |
| Personal data .....  | 262        |
| CIA principles .....   | 263        |
| Sample / Sampling.....   | 264        |
| Data processing / wrangling .....                                    | 264        |
| Data FAIRification.....  | 265        |
| Attributes .....   | 265        |
| Features.....  | 266        |
| Labels / data labels.....  | 267        |
| Proxies.....   | 267        |
| <b>B.8      Algorithm, model, algorithm-to-model transition.....</b> | <b>269</b> |
| Algorithm.....   | 269        |
| Machine learning.....  | 270        |
| ML versus statistics.....  | 271        |
| Machine learning algorithm.....                                      | 271        |
| Machine learning model .....   | 272        |
| Non-learning algorithm.....  | 273        |
| Objective function.....  | 274        |
| Parameters .....   | 276        |
| Hyperparameters .....  | 277        |
| Neural network bias .....  | 277        |
| Artificial neural networks.....                                      | 278        |
| Artificial neural networks – types.....                              | 280        |
| Deep learning .....  | 281        |
| Algorithm-to-model transition (ATMT) .....                           | 282        |
| Federated learning & split learning.....                             | 286        |
| Continuous and adaptive learning .....                               | 287        |
| Causal learning.....   | 288        |
| Causal machine learning.....   | 289        |
| Neurosymbolic AI .....   | 290        |
| Transfer learning.....   | 290        |
| Hybrid model / algorithm .....                                       | 292        |
| Hybrid learning .....  | 292        |
| Model development.....   | 293        |
| Evaluation metrics.....  | 294        |
| Model performance .....  | 296        |
| Model calibration.....   | 296        |
| Model actionability.....   | 297        |
| Causability.....   | 297        |
| Accuracy.....  | 298        |
| Sensitivity .....  | 299        |
| Specificity.....   | 299        |
| Precision.....   | 299        |
| Receiver operating characteristic (ROC) .....                        | 300        |
| Precision-recall (PR) curve.....                                     | 300        |
| Recall .....   | 300        |
| True positive rate .....   | 301        |
| Replicability.....   | 301        |
| Reproducibility .....  | 302        |
| Reliability .....  | 302        |
| Generalisability.....  | 303        |
| Overfitting .....  | 303        |
| Foundation models .....  | 304        |

|  |            |
|--|------------|
| Generative AI (GenAI) .....  | 305        |
| Embeddings .....   | 308        |
| <i>B.9 Relevance</i> .....   | 309        |
| Conceptual relevance .....   | 309        |
| Contextual relevance .....   | 310        |
| Valid clinical association / scientific validity .....                 | 310        |
| <i>B.10 Verification, validation, evaluation</i> .....                 | 312        |
| Verification .....   | 312        |
| Validation .....   | 312        |
| Model validation .....   | 313        |
| Model testing .....  | 316        |
| Model evaluation .....   | 316        |
| Model integrity checking .....   | 317        |
| AI system validation ('analytical / technical validation') .....       | 318        |
| Usability validation .....   | 319        |
| Clinical validation .....  | 321        |
| Continuous evaluation of AI systems .....                              | 324        |
| <i>B.11 Clinical concepts</i> .....                                    | 326        |
| Health outcomes .....  | 326        |
| Healthcare modality .....  | 326        |
| Efficacy, effectiveness and efficiency .....                           | 327        |
| Clinical effectiveness .....   | 328        |
| Clinical performance .....   | 328        |
| Clinical benefit .....   | 329        |
| Clinical safety .....  | 330        |
| Patient safety .....   | 330        |
| Clinical evidence .....  | 331        |
| Clinical data .....  | 331        |
| Clinical investigation .....   | 332        |
| Clinical endpoint .....  | 333        |
| Clinical investigation plan .....                                      | 333        |
| Clinical predictions .....   | 334        |
| Adverse event .....  | 335        |
| Incident .....   | 335        |
| Post-market surveillance, market surveillance, corrective action ..... | 336        |
| Post-market clinical follow-up (PMCF) .....                            | 337        |
| Personalised medicine & precision medicine .....                       | 338        |
| <i>B.12 Use of AI systems in health and healthcare</i> .....           | 340        |
| Use .....  | 340        |
| User research .....  | 341        |
| Usability .....  | 341        |
| User competency and training requirements .....                        | 342        |
| Intended use .....   | 342        |
| Foreseeable misuse .....   | 343        |
| Instructions for use .....   | 343        |
| Real-world use .....   | 344        |
| Use environment .....  | 344        |
| Use context .....  | 346        |
| Clinical practice guideline (CPG) .....                                | 347        |
| Clinical practice protocol .....                                       | 347        |
| Clinical pathway .....   | 348        |
| Applicability and limitations .....                                    | 348        |
| <b>5 Conclusion .....</b>  | <b>350</b> |
| <b>References .....</b>  | <b>351</b> |
| A .....  | 351        |
| B .....  | 353        |
| C .....  | 356        |
| Council of Europe .....  | 358        |

|  |            |
|--|------------|
| <i>D</i> .....   | 359        |
| <i>E</i> .....   | 361        |
| <i>European Commission</i> .....   | 362        |
| <i>European Union (EU)</i> .....   | 363        |
| <i>EU HLEG (EU high-level expert group)</i> .....                                    | 364        |
| <i>EU MDCG - medical devices coordination group</i> .....                            | 364        |
| <i>European Parliament</i> .....   | 365        |
| <i>F</i> .....   | 365        |
| <i>G</i> .....   | 367        |
| <i>H</i> .....   | 369        |
| <i>I</i> .....   | 371        |
| <i>IMDRF</i> .....   | 372        |
| <i>ISO</i> .....   | 373        |
| <i>J</i> .....   | 374        |
| <i>K</i> .....   | 374        |
| <i>L</i> .....   | 376        |
| <i>M</i> .....   | 378        |
| <i>N</i> .....   | 381        |
| <i>NIST</i> .....  | 382        |
| <i>O</i> .....   | 383        |
| <i>OECD</i> .....  | 383        |
| <i>P</i> .....   | 384        |
| <i>Q</i> .....   | 386        |
| <i>R</i> .....   | 386        |
| <i>S</i> .....   | 388        |
| <i>T</i> .....   | 393        |
| <i>U</i> .....   | 394        |
| <i>United Nations</i> .....  | 394        |
| <i>US FDA</i> .....  | 395        |
| <i>V</i> .....   | 396        |
| <i>W</i> .....   | 397        |
| <i>World Economic Forum</i> .....  | 398        |
| <i>World Health Organisation (WHO)</i> .....   | 398        |
| <i>X</i> .....   | 400        |
| <i>Y</i> .....   | 400        |
| <i>Z</i> .....   | 400        |
| <b>List of abbreviations and definitions</b> .....                                   | <b>402</b> |
| <b>List of boxes</b> .....   | <b>404</b> |
| <b>List of figures</b> .....   | <b>405</b> |
| <b>List of tables</b> .....  | <b>409</b> |
| <b>Annexes</b> .....   | <b>410</b> |
| <i>Annex 1 – Tools for clinical studies and evaluation of AI in healthcare</i> ..... | 410        |
| EQUATOR network: enhancing the quality and transparency of health research .....     | 410        |
| STARD-AI .....   | 410        |
| TRIPOD-AI .....  | 411        |
| PROBAST-AI .....   | 411        |
| DECIDE-AI .....  | 412        |
| SPIRIT-AI .....  | 412        |
| CONSORT-AI .....   | 413        |
| CLAIM .....  | 413        |
| MI-CLAIM .....   | 414        |
| FUTURE-AI .....  | 415        |
| QUADAS-AI .....  | 415        |
| CORE-MD clinical risk score .....  | 416        |
| <i>Annex 2 - Detailed table of contents</i> .....                                    | 417        |
| <i>Annex 3 - Alphabetical list of entities of this ontology</i> .....                | 425        |

## **Annex 3 – Alphabetical list of entities of this ontology**

|   |     |
|---|-----|
| A priori trust in AI systems.....   | 51  |
| Accessibility of data and infrastructure for AI development.....              | 166 |
| Accountability structures, attribution of (distributed) responsibilities..... | 143 |
| Accountability.....   | 143 |
| Accuracy gain.....  | 59  |
| Accuracy.....   | 298 |
| Acting with integrity.....  | 141 |
| Actors as defined in the EU's AI Act.....                                     | 217 |
| Added value.....  | 57  |
| Adverse event.....  | 335 |
| Affective computing.....  | 224 |
| Agency.....   | 222 |
| AI actors.....  | 214 |
| AI and the development of healthcare.....                                     | 87  |
| AI as a scientific field.....   | 175 |
| AI ethics.....  | 191 |
| AI Evidence pathway for health.....   | 241 |
| AI governance.....  | 201 |
| AI impact assessment (AI-IA) .....  | 197 |
| AI management.....  | 204 |
| AI practitioners.....   | 215 |
| AI principles and AI ethics guidelines .....                                  | 192 |
| AI risk management.....   | 205 |
| AI safety.....  | 208 |
| AI system component / part.....   | 179 |
| AI system output .....  | 182 |
| AI system validation ('analytical / technical validation') .....              | 318 |
| AI systems .....  | 178 |
| AI technique.....   | 184 |
| AI typology.....  | 183 |
| AI value chain actors.....  | 219 |
| AI-enabled medical device software.....                                       | 180 |
| AI-enabled software as a medical device (AI-SAMD).....                        | 182 |
| AI-enabled software in a medical device (AI-SIMD).....                        | 182 |
| Algorithm.....  | 269 |
| Algorithmic bias .....  | 237 |
| Algorithm-to-model transition (ATMT).....                                     | 282 |
| Alignment.....  | 199 |
| Applicability and limitations .....   | 348 |
| Artificial intelligence (AI).....   | 174 |
| Artificial neural networks – types.....                                       | 280 |
| Artificial neural networks.....   | 278 |
| Attributes.....   | 265 |
| Auditability and auditing.....  | 144 |
| Augmentation.....   | 229 |
| Automation.....   | 229 |
| Autonomy .....  | 225 |

|  |     |
|--|-----|
| Avoiding automation bias .....   | 90  |
| Avoiding automation complacency .....  | 91  |
| Avoiding colonial structures; reducing the global north-south divide .....       | 166 |
| Avoiding discrimination and discriminatory bias .....                            | 156 |
| Beneficence: translational concepts .....  | 56  |
| Bias .....   | 231 |
| Bioethics .....  | 189 |
| Causability .....  | 297 |
| Causal learning .....  | 288 |
| Causal machine learning .....  | 289 |
| CIA principle from a privacy protection perspective .....                        | 105 |
| CIA principles .....   | 263 |
| Clinical benefit .....   | 329 |
| Clinical benefits for patient .....  | 61  |
| Clinical data .....  | 331 |
| Clinical effectiveness .....   | 328 |
| Clinical endpoint .....  | 333 |
| Clinical evidence .....  | 331 |
| Clinical gains .....   | 61  |
| Clinical investigation plan .....  | 333 |
| Clinical investigation .....   | 332 |
| Clinical pathway .....   | 348 |
| Clinical performance .....   | 328 |
| Clinical practice guideline (CPG) .....  | 347 |
| Clinical practice protocol .....   | 347 |
| Clinical predictions .....   | 334 |
| Clinical safety .....  | 330 |
| Clinical validation .....  | 321 |
| Communication .....  | 112 |
| Compliance experts .....   | 220 |
| Computational theory of mind (CTM) .....   | 176 |
| Concept drift / shift .....  | 239 |
| Conceptual relevance .....   | 309 |
| Conditions for trust and trustworthiness: actors and AI systems .....            | 50  |
| Consent concerning collection of personal data / medical information .....       | 95  |
| Considering mental health impacts of AI exposure .....                           | 166 |
| Considering patients with rare diseases / conditions .....                       | 166 |
| Consistency gain .....   | 59  |
| Contestability and challenge .....   | 146 |
| Contextual relevance .....   | 310 |
| Continuous and adaptive learning .....   | 287 |
| Continuous evaluation of AI systems .....  | 324 |
| Conversational agents .....  | 185 |
| Correcting problems and failures, including necessary communication .....        | 142 |
| Corrigibility .....  | 227 |
| Cybersecurity: risks for value chain assets and knock-on effects on safety ..... | 77  |
| Data access or collection, wrangling .....                                       | 64  |
| Data drift / shift .....   | 239 |
| Data FAIRification .....   | 265 |
| Data governance, data management & data accountability .....                     | 104 |
| Data modality .....  | 259 |

|   |     |
|---|-----|
| Data privacy .....  | 261 |
| Data processing / wrangling.....  | 264 |
| Data protection.....  | 101 |
| Data provenance of development data .....                                 | 252 |
| Data quality metrics.....   | 257 |
| Data quality .....  | 255 |
| Data security.....  | 105 |
| Dataset.....  | 252 |
| Datasheets for datasets.....  | 260 |
| Decommissioning / Retirement.....   | 250 |
| Decommissioning.....  | 65  |
| Deep learning.....  | 281 |
| Deskilling .....  | 91  |
| Development data .....  | 252 |
| Digital infrastructure gains .....  | 64  |
| Dignity: translational concepts .....                                     | 87  |
| Disclosure.....   | 113 |
| Distributional drift / shift.....   | 240 |
| Drift / shift in machine learning.....                                    | 238 |
| Economic gain - person time .....   | 60  |
| Economic gain: investments.....   | 60  |
| Economic gain: less need for non-human resources.....                     | 60  |
| Economic gain: returns .....  | 60  |
| Economic gains.....   | 60  |
| Efficacy, effectiveness and efficiency .....                              | 327 |
| Embeddings .....  | 308 |
| Ensuring human agency and oversight.....                                  | 144 |
| Ensuring the means for free and informed consent.....                     | 92  |
| Environmental gain.....   | 65  |
| Environmental sustainability of AI system throughout the life cycle ..... | 173 |
| Ethical evaluation of AI .....  | 195 |
| Ethical principles .....  | 187 |
| Ethics code & governance / management frameworks.....                     | 141 |
| Ethics code .....   | 200 |
| Evaluation metrics.....   | 294 |
| Evidence about frameworks and processes used .....                        | 116 |
| Evidence on algorithm-to-model transition.....                            | 116 |
| Evidence on data .....  | 116 |
| Evidence on intended use, applicability and limitations .....             | 117 |
| Evidence on interoperability and value chain elements .....               | 118 |
| Explanations of AI systems and their outcomes .....                       | 131 |
| Explicability .....   | 129 |
| Failure transparency.....   | 114 |
| Fairness: translational concepts .....                                    | 156 |
| Features .....  | 266 |
| Federated learning & split learning.....                                  | 286 |
| Foreseeable misuse .....  | 343 |
| Foundation models.....  | 304 |
| Fundamental rights and algorithm impact assessments .....                 | 198 |
| Gain in regard to complex ethical considerations in healthcare .....      | 63  |
| Gains for population / public health .....                                | 63  |

|   |     |
|---|-----|
| Gains .....   | 59  |
| General gains of operational efficiency and usability .....   | 60  |
| General health system management and planning gain.....   | 64  |
| General healthcare management gains .....   | 62  |
| General output quality gains.....   | 59  |
| Generalisability.....   | 303 |
| Generalisability: shortcut learning .....   | 76  |
| Generating and evaluating evidence on risks .....   | 71  |
| Generating and evaluation evidence on benefits .....  | 56  |
| Generating evidence on wider impacts .....  | 83  |
| Generative AI (GenAI).....  | 305 |
| Global healthcare equality & equity gain(s).....  | 62  |
| Health equality & health equity.....  | 159 |
| Health outcomes.....  | 326 |
| Health policy makers, health economists, health system managers.....  | 221 |
| Health research gains.....  | 65  |
| Health technology assessment (HTA) experts .....  | 220 |
| Health technology assessment .....  | 248 |
| Healthcare accessibility gains.....   | 62  |
| Healthcare adaptation gains.....  | 62  |
| Healthcare efficiency gain.....   | 62  |
| Healthcare equality gains .....   | 62  |
| Healthcare equity gains .....   | 62  |
| Healthcare gain - personalisation of prevention/treatment.....  | 63  |
| Healthcare gain - unmet medical or patient needs.....   | 63  |
| Healthcare gains .....  | 62  |
| Healthcare modality .....   | 326 |
| Healthcare precision gain.....  | 62  |
| Healthcare professional well-being.....   | 61  |
| Healthcare quality gain .....   | 62  |
| Heuristics .....  | 234 |
| Higher degree of autonomy / independence of patients with chronic diseases and/or debilitating conditions ..... | 61  |
| Human agency .....  | 223 |
| Human oversight .....   | 224 |
| Human primacy .....   | 226 |
| Human-centric AI .....  | 211 |
| Hybrid learning .....   | 292 |
| Hybrid model / algorithm .....  | 292 |
| Hyperparameters.....  | 277 |
| Impact assessments .....  | 83  |
| Impact assessments: risk identification, data protection, benefits .....  | 102 |
| Improved patient-physician relationship.....  | 61  |
| Incident .....  | 335 |
| Inclusivity gain .....  | 63  |
| Information on training requirements.....   | 119 |
| Input data.....   | 254 |
| Instructions for use .....  | 343 |
| Intelligibility .....   | 119 |
| Intended use .....  | 342 |
| Interpretability and explainability .....   | 124 |

|  |     |
|--|-----|
| Intrinsic incompatibilities or ‘trade-offs’ .....  | 237 |
| Job impacts.....   | 167 |
| Job impacts, skills, training.....   | 167 |
| Labels / data labels.....  | 267 |
| Lawful, legitimate and fair use of data.....   | 101 |
| Legal and regulatory compliance.....   | 141 |
| Liability .....  | 147 |
| Life cycle of AI in health.....  | 243 |
| Looking out for and avoiding discriminatory bias.....                                      | 157 |
| Machine agency.....  | 228 |
| Machine learning algorithm.....  | 271 |
| Machine learning model.....  | 272 |
| Machine learning .....   | 270 |
| Management of right of access (e.g. EU’s GDPR, Data Act).....                              | 102 |
| Medical privacy / health privacy.....  | 95  |
| ML versus statistics .....   | 271 |
| Model actionability .....  | 297 |
| Model calibration.....   | 296 |
| Model development .....  | 293 |
| Model evaluation .....   | 316 |
| Model integrity checking.....  | 317 |
| Model performance.....   | 296 |
| Model testing .....  | 316 |
| Model validation.....  | 313 |
| Monitoring and mitigation of possible discrimination throughout the evidence pathway ..... | 158 |
| Monitoring drifts / shifts .....   | 83  |
| Monitoring incidents / adverse events .....  | 82  |
| Monitoring performance, effectiveness, efficiency and bias.....                            | 83  |
| Monitoring usability and usability-related errors.....                                     | 83  |
| Nature of data .....   | 260 |
| Neural network bias.....   | 277 |
| Neurosymbolic AI .....   | 290 |
| Non-bias gain & equitable bias gains.....  | 59  |
| Non-learning algorithm .....   | 273 |
| Non-maleficence: translational concepts.....   | 71  |
| Objective function.....  | 274 |
| Open, democratic and evidence-based debate .....   | 88  |
| Open, democratic and solidary society, culture of dialogue .....                           | 165 |
| Output and output data .....   | 255 |
| Overfitting.....   | 303 |
| Parameters.....  | 276 |
| Patient primacy .....  | 228 |
| Patient safety .....   | 330 |
| Patient well-being.....  | 61  |
| Peer review and community discourse.....   | 142 |
| Personal data.....   | 262 |
| Personalised medicine & precision medicine .....   | 338 |
| Post-deployment input data.....  | 259 |
| Post-deployment monitoring .....   | 249 |
| Post-market clinical follow-up (PMCF) .....  | 337 |
| Post-market surveillance, market surveillance, corrective action.....                      | 336 |

|  |     |
|--|-----|
| Potential benefits.....  | 59  |
| Precision gain.....  | 59  |
| Precision.....   | 299 |
| Precision-recall (PR) curve .....  | 300 |
| Prevention of unauthorised access (e.g. servers, devices, platforms, cloud) .....                    | 106 |
| Principlism.....   | 190 |
| Privacy protection: translational concepts.....  | 101 |
| Privacy-preserving techniques for AI development.....  | 107 |
| Private-public partnerships.....   | 169 |
| Proxies.....   | 267 |
| Quality culture and risk management .....  | 142 |
| Real-world benefits.....   | 58  |
| Real-world use .....   | 344 |
| Recall.....  | 300 |
| Receiver operating characteristic (ROC).....   | 300 |
| Reduced burden for HCP.....  | 60  |
| Reducing heuristic decision-making.....  | 61  |
| Reliability.....   | 302 |
| Remedy and redress.....  | 147 |
| Replicability.....   | 301 |
| Reproducibility.....   | 302 |
| Research involving human subjects: ethical principles & guidelines.....                              | 212 |
| Respecting patient primacy .....   | 87  |
| Responsibility: translational concepts .....   | 141 |
| Responsiveness .....   | 146 |
| Right of intelligible information about personal data being processed .....                          | 96  |
| Right to demand rectification or erasure of data processed not in line with relevant provisions..... | 97  |
| Right to know and right not to know.....   | 94  |
| Right to know if AI system employed.....   | 94  |
| Risk.....  | 207 |
| Risks related to bias .....  | 74  |
| Risks related to data privacy of personal information.....   | 72  |
| Risks related to dignity, freedom and autonomy.....  | 71  |
| Risks related to insufficient post-deployment monitoring / post-market surveillance.....             | 82  |
| Risks related to insufficient robustness / resilience and generalisability.....                      | 75  |
| Risks related to integration into workflows.....   | 81  |
| Risks related to pre-deployment evidence gaps .....  | 82  |
| Risks related to responsibility.....   | 73  |
| Risks related to transparency .....  | 73  |
| Risks related to value chain elements .....  | 78  |
| Risks relating to technical integration and interoperability issues .....                            | 80  |
| Robustness / resilience: use context and use environment variations & drift issues .....             | 75  |
| Safe innovation: community bridging, collaboration, education .....                                  | 168 |
| Sample / Sampling .....  | 264 |
| Sensitivity.....   | 299 |
| Shortcut learning.....   | 236 |
| Social and societal gains, including public health .....   | 63  |
| Societal gain.....   | 63  |
| Socioeconomic gain .....   | 64  |
| Solidarity: translational concepts .....   | 165 |
| Specific aspects of trust in the health domain .....   | 51  |

|   |     |
|---|-----|
| Specificity.....  | 299 |
| Sustainability of health systems using AI .....                               | 173 |
| Sustainability: translational concepts.....                                   | 173 |
| Synthetic data .....  | 64  |
| Synthetic health data .....   | 260 |
| Taking vulnerable groups into account “by design”.....                        | 165 |
| Testing data.....   | 254 |
| Time-to-delivery gain .....   | 60  |
| Traceability.....   | 114 |
| Training and skills .....   | 167 |
| Training data.....  | 253 |
| Transfer learning.....  | 290 |
| Transparency gain .....   | 64  |
| Transparency of AI systems.....   | 115 |
| Transparency of human-AI interaction.....                                     | 113 |
| Transparency of organisations and actors providing/deploying AI systems ..... | 112 |
| Transparency: translational concepts.....                                     | 112 |
| True positive rate .....  | 301 |
| Trust and trustworthiness: translational concepts.....                        | 49  |
| Trust issues related to situations of conflict of interest (CoI).....         | 51  |
| Unavoidable trade-offs .....  | 160 |
| Understandable explanations.....  | 137 |
| Universal versus targeted design .....  | 161 |
| Upholding a trustful patient-physician relationship.....                      | 89  |
| Usability gain .....  | 60  |
| Usability validation.....   | 319 |
| Usability .....   | 341 |
| Use context .....   | 346 |
| Use environment.....  | 344 |
| Use.....  | 340 |
| User competency and training requirements .....                               | 342 |
| User research.....  | 341 |
| Users of AI in the health domain .....  | 217 |
| Valid clinical association / scientific validity.....                         | 310 |
| Validation data .....   | 253 |
| Validation .....  | 312 |
| Value chain of AI .....   | 246 |
| Verification.....   | 312 |
| Well-being gains .....  | 61  |

## **Getting in touch with the EU**

### **In person**

All over the European Union there are hundreds of Europe Direct centres. You can find the address of the centre nearest you online ([european-union.europa.eu/contact-eu/meet-us\\_en](http://european-union.europa.eu/contact-eu/meet-us_en)).

### **On the phone or in writing**

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
- at the following standard number: +32 22999696,
- via the following form: [european-union.europa.eu/contact-eu/write-us\\_en](http://european-union.europa.eu/contact-eu/write-us_en).

## **Finding information about the EU**

### **Online**

Information about the European Union in all the official languages of the EU is available on the Europa website ([european-union.europa.eu](http://european-union.europa.eu)).

### **EU publications**

You can view or order EU publications at [op.europa.eu/en/publications](http://op.europa.eu/en/publications). Multiple copies of free publications can be obtained by contacting Europe Direct or your local documentation centre ([european-union.europa.eu/contact-eu/meet-us\\_en](http://european-union.europa.eu/contact-eu/meet-us_en)).

### **EU law and related documents**

For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex ([eur-lex.europa.eu](http://eur-lex.europa.eu)).

### **EU open data**

The portal [data.europa.eu](http://data.europa.eu) provides access to open datasets from the EU institutions, bodies and agencies. These can be downloaded and reused for free, for both commercial and non-commercial purposes. The portal also provides access to a wealth of datasets from European countries.

# Science for policy

The Joint Research Centre (JRC) provides independent, evidence-based knowledge and science, supporting EU policies to positively impact society



**EU Science Hub**

[Joint-research-centre.ec.europa.eu](http://Joint-research-centre.ec.europa.eu)



Publications Office  
of the European Union