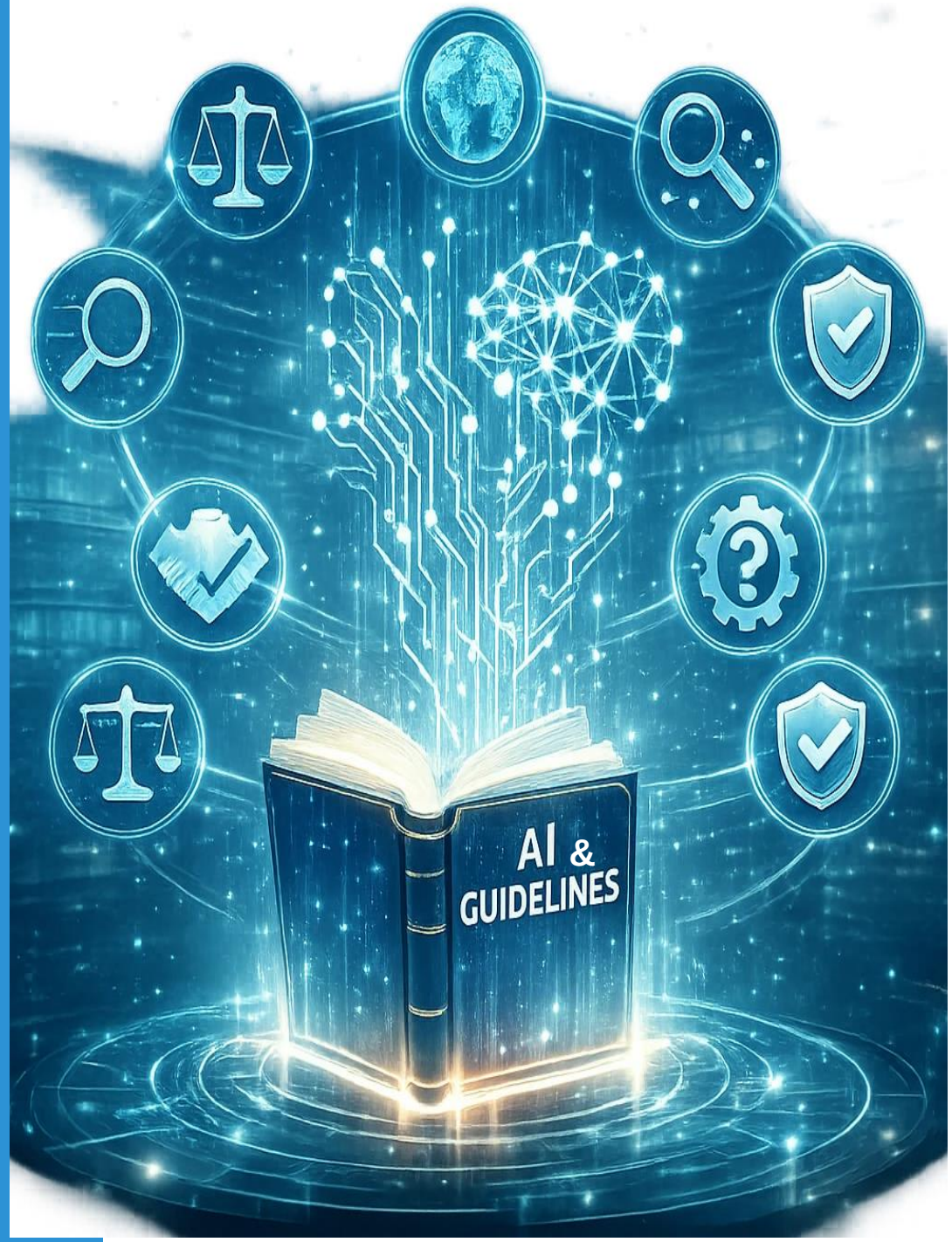


FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare

Adviser: Dr Aynaz Nourani

Elahe Hosseini

Item	Description
Journal Name	BMJ (British Medical Journal)
Publisher	BMJ Publishing Group Ltd
Country	United Kingdom
Indexing Databases	Web of Science (ISI–SCI-E), Scopus, PubMed/MEDLINE
Quartile (Q)	Q1
Subject Category	Medicine, General & Internal
Journal Rank (Percentile)	Above 99% of journals in the same category
Impact Factor (IF)	42.7
Accepted	10 January 2025


Introduction

Introduction

- A clear gap persists between building innovative AI models and their integration into clinical practice.
- Despite major advances in AI research for healthcare, real-world clinical adoption remains limited.
- This gap is driven by multiple factors, including risks of error, data bias, and limited transparency.

Introduction

- Healthcare AI faces technical, clinical, ethical, and societal risks.
- Existing AI tools can be prone to errors, patient harm, bias, and health inequalities.
- Lack of transparency, accountability, data privacy and security are common concerns.

- Healthcare has unique trust dynamics between doctors and patients, requiring robust and transparent AI.
- AI lacks universally accepted quality assurance measures, unlike medical equipment.
- Errors in healthcare AI can have major consequences compared with other domains.

Introduction

- Existing guidelines (TRIPOD-AI, CLAIM, CONSORT-AI) focus mainly on reporting standards.
- They lack best practices for development and deployment of AI tools.
- Most proposals are not based on wide international consensus or full AI lifecycle coverage.

- FUTURE-AI is the first structured, holistic guideline for trustworthy AI in healthcare.
- The FUTURE-AI Consortium was Established in 2021 by international consensus among 117 experts from 50 countries.
- Framework covers six guiding principles: fairness, universality, traceability, usability, robustness, and explainability.
- 30 best practices address technical, clinical, socioethical, and legal aspects across the AI lifecycle.

Introduction

How did the practical activities begin?

- The work started within the AI4HI (AI for Health Imaging) network
- This network consisted of several European projects:
  - EuCanImage
  - ProCancer-I
  - CHAIMELEON
  - PRIMAGE
  - INCISIVE
- The initial focus was on artificial intelligence in medical imaging
- Subsequently, in Round 3, the scope was broadened to AI across the entire healthcare domain

Introduction

- The FUTURE-AI framework incorporates multidisciplinary input from experts worldwide.
- It covers all continents and several disciplines (data science, medicine, ethics, social sciences).
- Designed as a dynamic framework, evolving with technological advances and stakeholder feedback.


Methodology

Overall Methodology

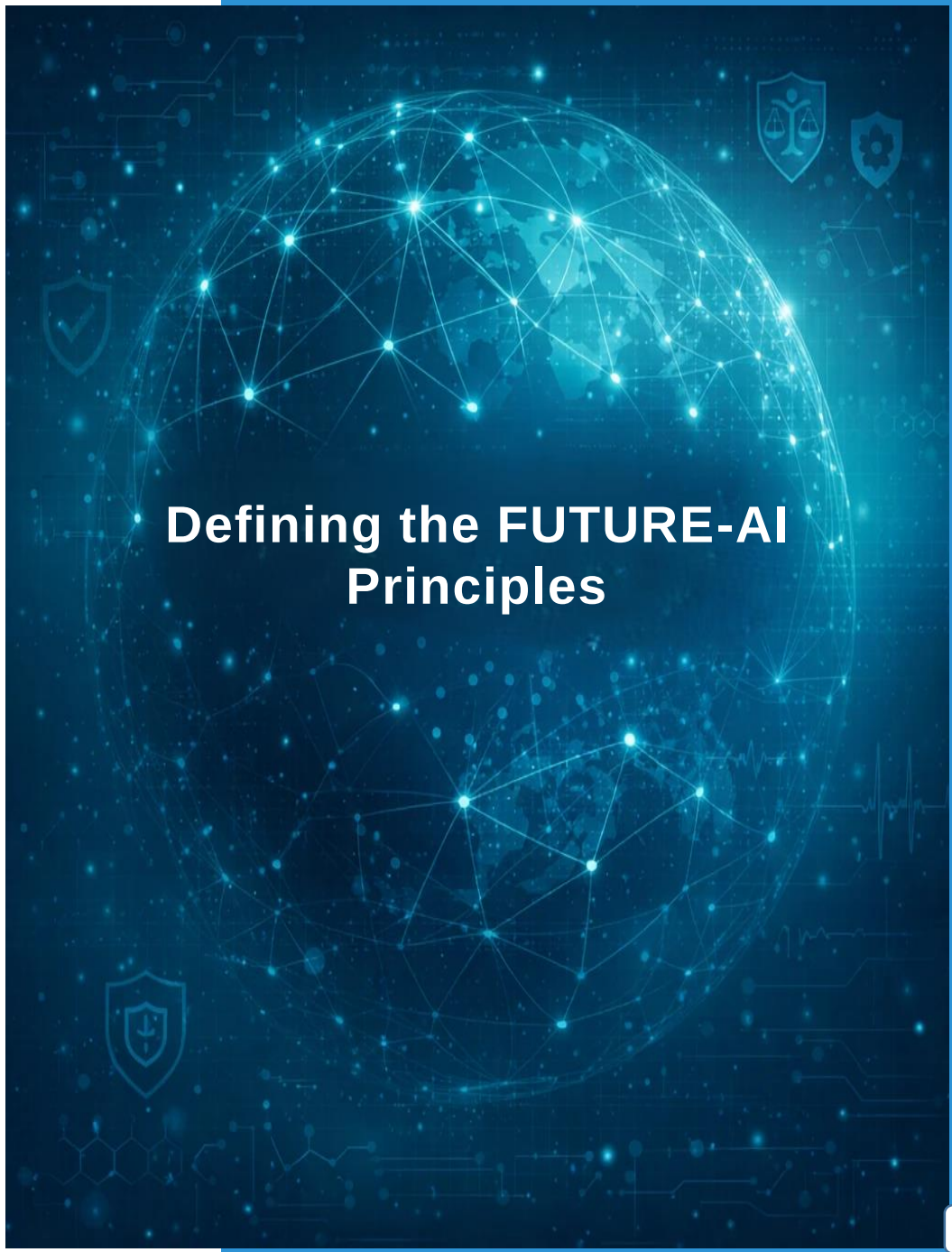
- FUTURE-AI is a structured framework providing guiding principles and step-by-step recommendations for trustworthy healthcare AI
- FUTURE-AI guideline developed through international consensus over 24 months
- Methodology based on a Modified Delphi approach
- Eight iterative rounds of feedback and discussion
- Combination of quantitative voting and qualitative thematic analysis

Delphi approach

Overall Methodology:

- Inspired by the **FAIR** data management principles
- Aim: minimal, memorable, and user-friendly structure
- Extensive literature review on **trustworthy** and **ethical AI**
- **Six guiding principles** identified and clustered
- Additional searches included:

  Guidelines and position statements from major public bodies such as the **European Union (EU), US Food and Drug Administration (FDA), and WHO**

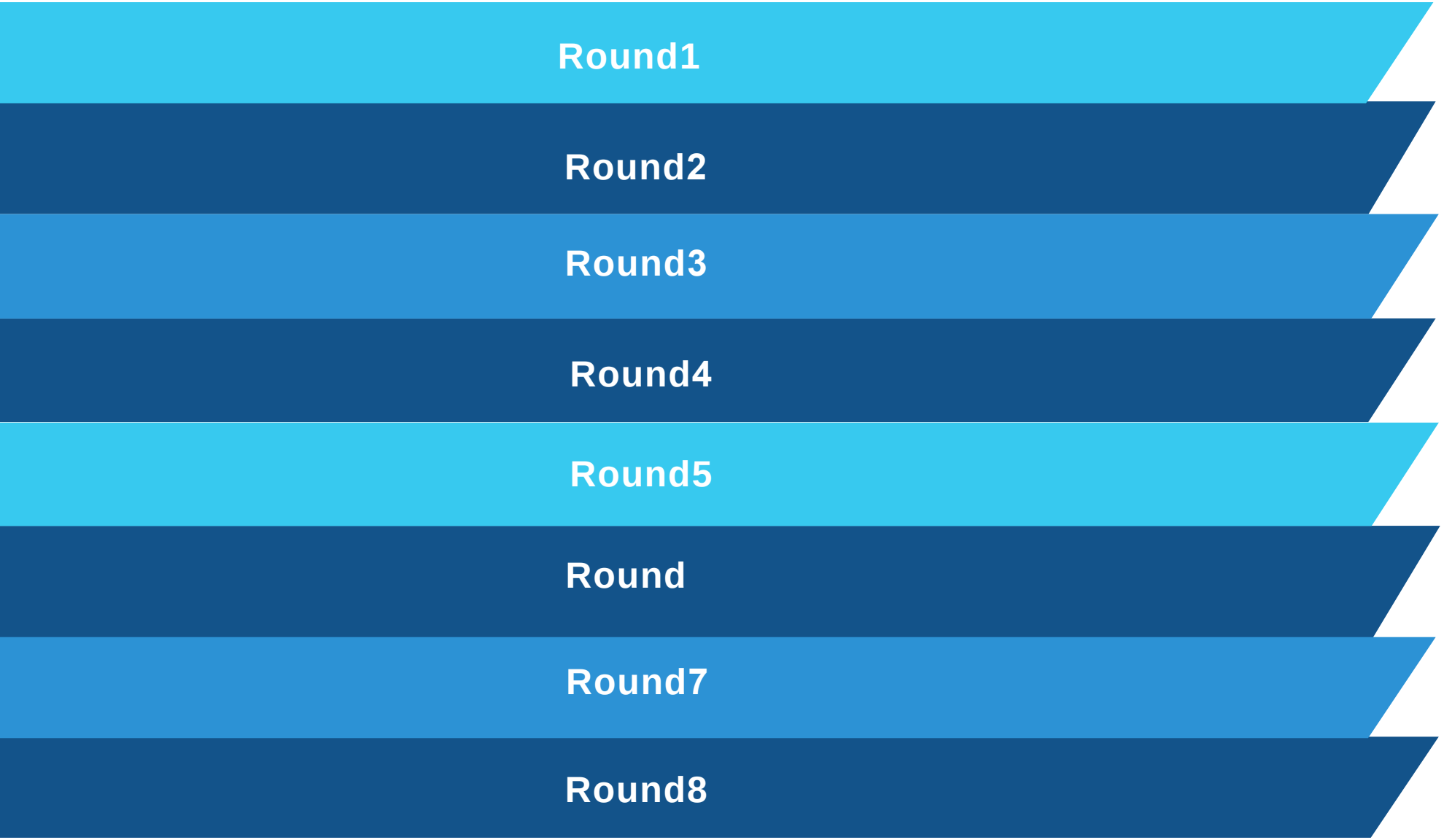


Overall Methodology

Clusters of requirements	Core principles
1. Fairness, diversity, inclusiveness, non-discrimination, unbiased AI, equity	Fairness
2. Generalisability, adaptability, interoperability, applicability, universality	Universality
3. Traceability, monitoring, continuous learning, auditing, accountability	Traceability
4. Human centred AI, user engagement, usability, accessibility, efficiency	Usability
5. Robustness, reliability, resilience, safety, security	Robustness
6. Transparency, explainability, interpretability, understandability	Explainability

Table 1 | Clustering of trustworthy artificial intelligence (AI) requirements and selection of FUTURE-AI guiding principles

Eight iterative rounds



Eight iterative rounds

Round 1

- Six working groups, each focusing on one principle
- Experts from AI4HI imaging projects
- Common use case: AI in medical imaging
- Output: 55 preliminary recommendations

Round2

- Consortium expanded to 72 experts
- Online survey with structured voting
- Consensus threshold: >90% agreement
- Output: 22 recommendations + 16 contentious points
- Introduction of a “General” category

Round3

- Feedback on adequacy and wording
- Resolution of several contentious issues
- Scope expanded from medical imaging to healthcare
- Guideline expanded to 30 best practices

Round4

- Focus on clarity, feasibility, and relevance
- Refinement of ambiguous terms
- Distinction between research and deployable AI
- Introduction of compliance levels(+/++)

Eight iterative rounds

Round5

- First full FUTURE-AI manuscript drafted by first and last authors
- Draft circulated among experts for iterative feedback
- Aim: ensure precision, clarity, and consistency of the guideline
- Integration of diverse perspectives (clinical, technical, non-technical)
- Experts suggested additional resources and references
- Practical method examples added to illustrate real-world implementation

Round6

- 44 new external experts invited for independent review
- Included patient advocates, social scientists, and regulatory experts
- Ensured broader geographical diversity (Africa, Latin America, Asia)
- Written feedback and structured voting (agree / disagree / neutral / unclear / no opinion)
- Main issues identified: misinterpretation and lack of clarity
- Helped pinpoint remaining ambiguities and formulations needing refinement

Round7

- Remaining contentious topics identified from previous rounds
- Four online consensus meetings held (June 2023)
- Focus on exact wording of recommendations
- Final wording and list of recommendations established

Round8

- Final vote conducted via an online survey
- Final consortium included 117 experts
- Original 72 experts from round 2
- Part of the 44 external experts from round 6
- Additional recruited experts
- All FUTURE-AI recommendations approved
- Less than 5% disagreement across all recommendations

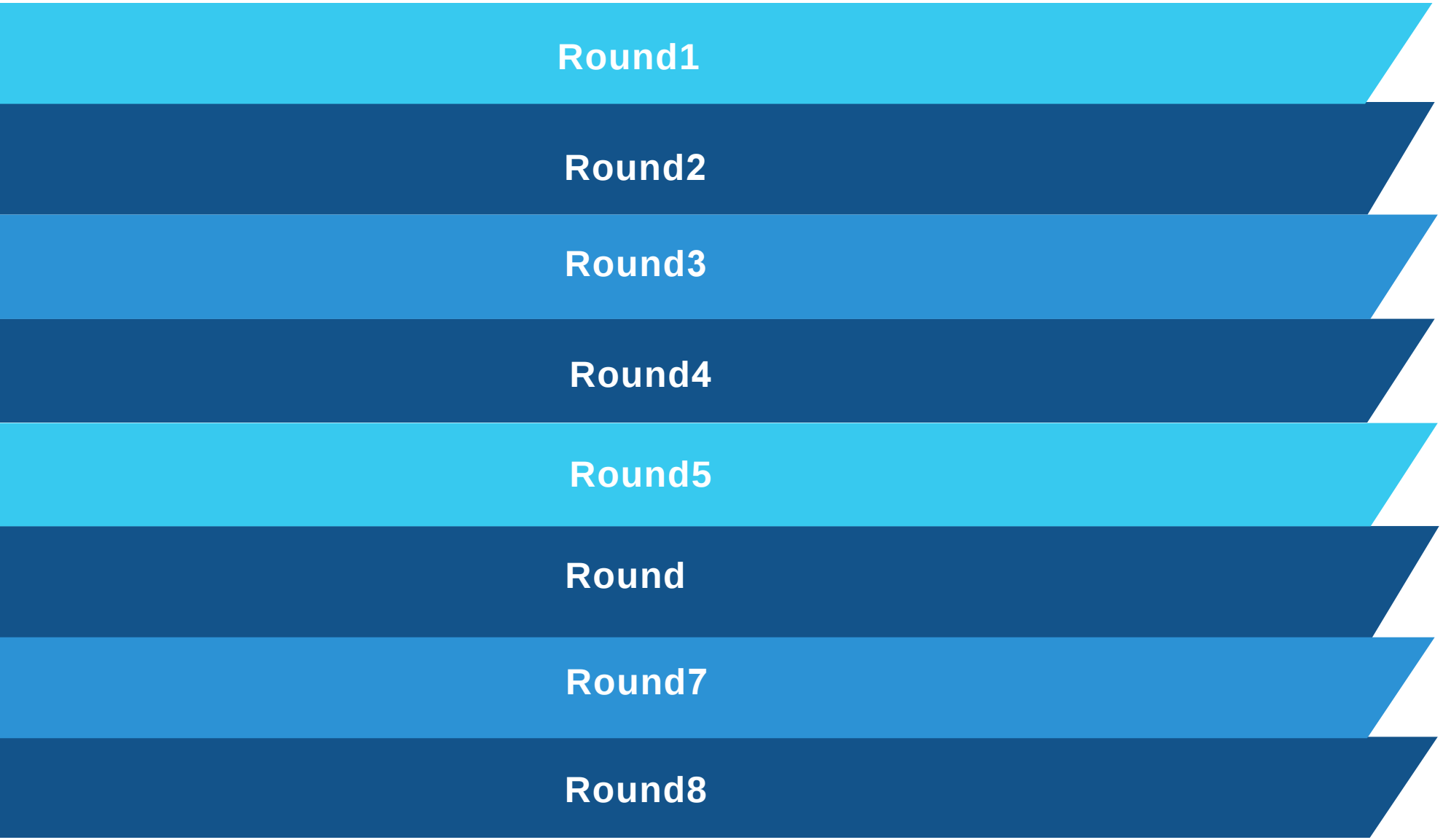
Clarifying the distinction between “Recommended” and “Highly recommended”

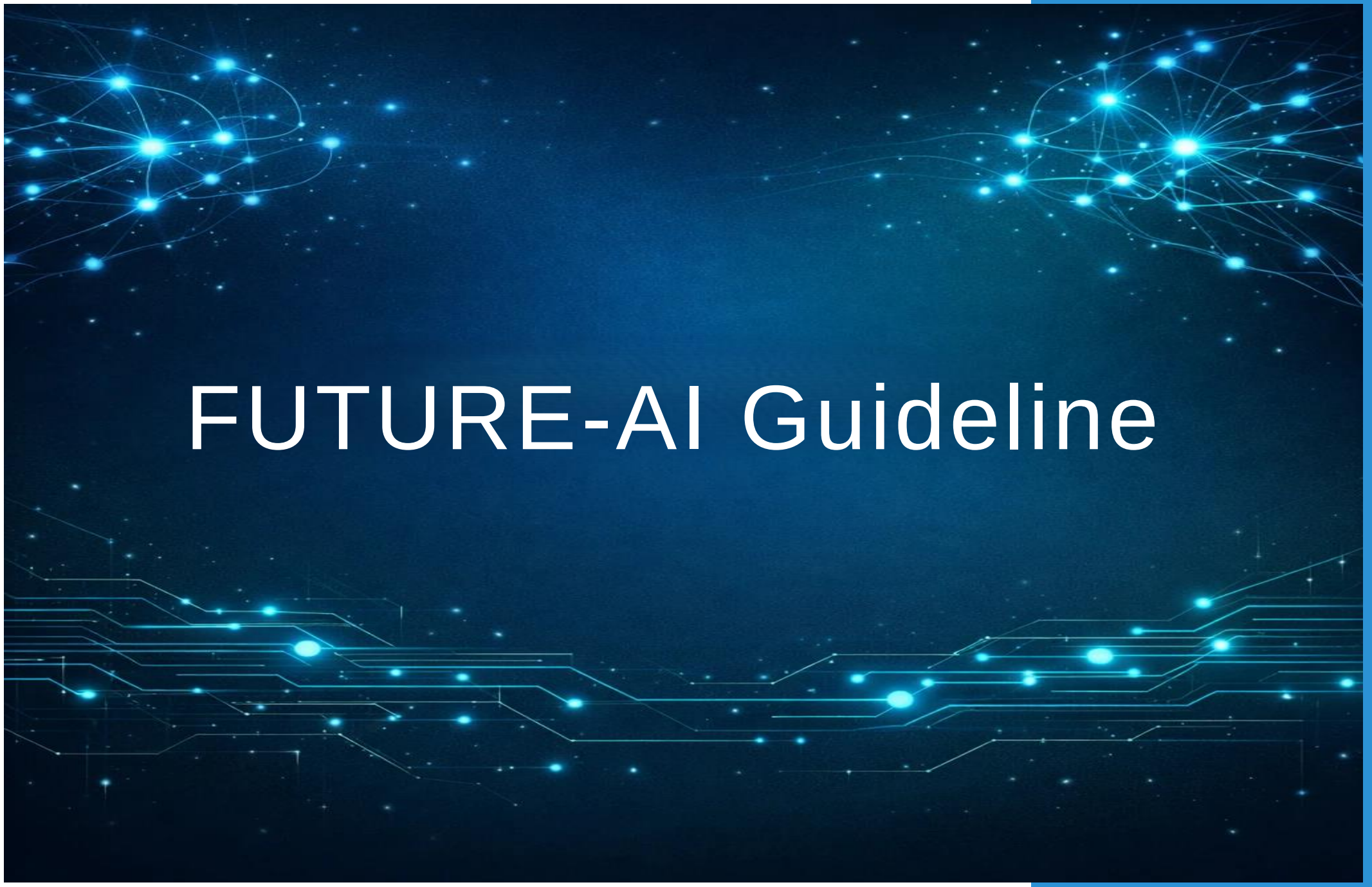
Table 2 | List of FUTURE-AI recommendations, together with the expected compliance for both research and deployable artificial intelligence (AI) tools

(+ : recommended ‘++ : highly recommended)

Category	Recommendations	Research	Deployable
Fairness	1. Define any potential sources of bias from an early stage	++	++
	2. Collect information on individuals’ and data attributes	+	+
	3. Evaluate potential biases and, when needed, bias correction measures	+	++
Universality	1. Define intended clinical settings and cross-setting variations	++	++
	2. Use community-defined standards (eg, clinical definitions, technical standards)	+	+
	3. Evaluate using external datasets and/or multiple sites	++	++
	4. Evaluate and demonstrate local clinical validity	+	++
Traceability	1. Implement a risk management process throughout the AI lifecycle	+	++
	2. Provide documentation (eg, technical, clinical)	++	++
	3. Define mechanisms for quality control of the AI inputs and outputs	+	++
	4. Implement a system for periodic auditing and updating	+	++
	5. Implement a logging system for usage recording	+	++
	6. Establish mechanisms for AI governance	+	++
Usability	1. Define intended use and user requirements from an early stage	++	++
	2. Establish mechanisms for human-AI interactions and oversight	+	++
	3. Provide training materials and activities (eg, tutorials, hands-on sessions)	+	++
	4. Evaluate user experience and acceptance with independent end users	+	++
	5. Evaluate clinical utility and safety (eg, effectiveness, harm, cost-benefit)	+	++
Robustness	1. Define sources of data variation from an early stage	++	++
	2. Train with representative real-world data	++	++
	3. Evaluate and optimise robustness against real-world variations	++	++
Explainability	1. Define the need and requirements for explainability with end users	++	++
	2. Evaluate explainability with end users (eg, correctness, impact on users)	+	+
General	1. Engage interdisciplinary stakeholders throughout the AI lifecycle	++	++
	2. Implement measures for data privacy and security	++	++
	3. Implement measures to address identified AI risks	++	++
	4. Define adequate evaluation plan (eg, datasets, metrics, reference methods)	++	++
	5. Identify and comply with applicable AI regulatory requirements	+	++
	6. Investigate and address application-specific ethical issues	+	++
	7. Investigate and address social and societal issues	+	+

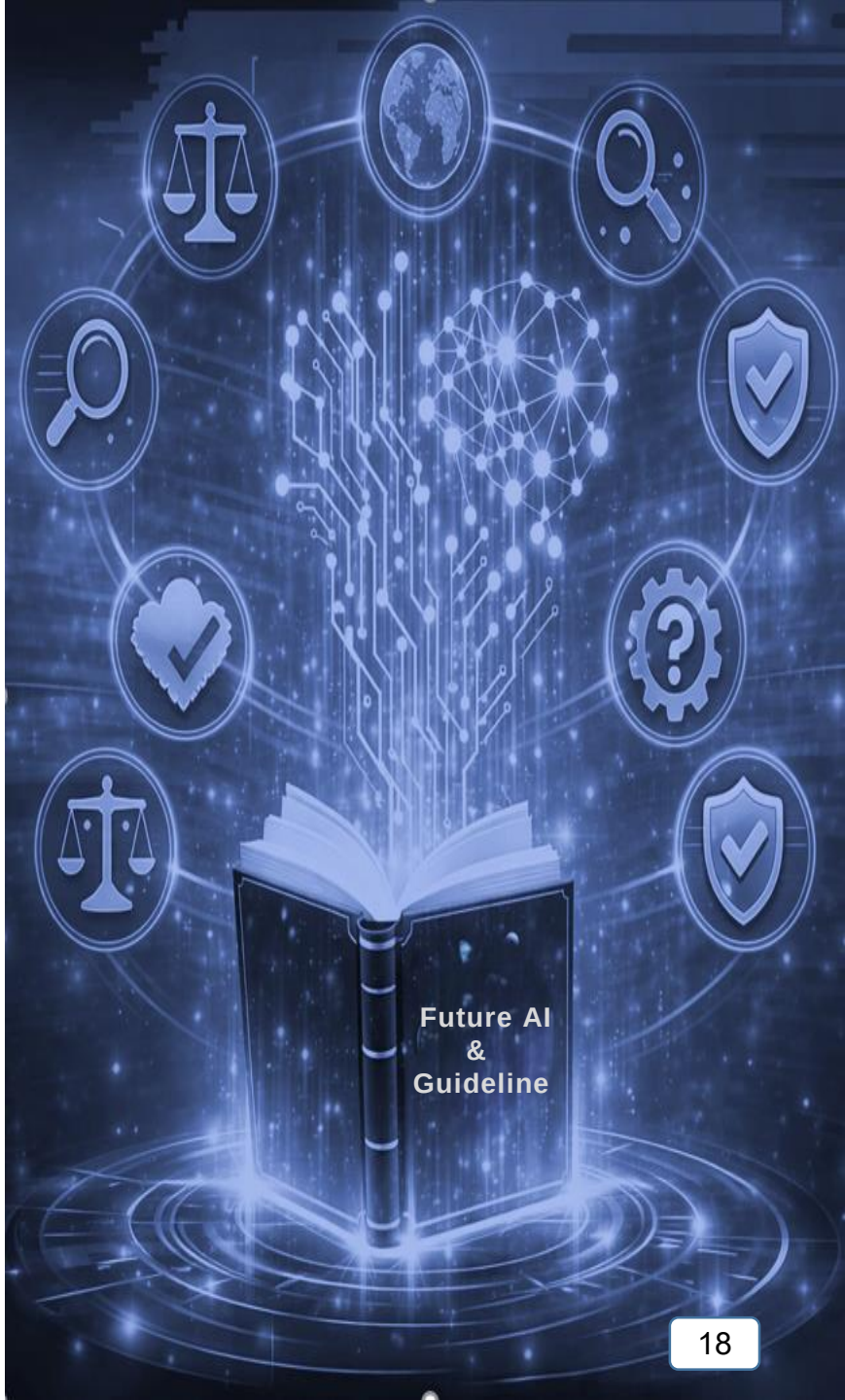
Eight iterative rounds




FUTURE-AI Guideline

FUTURE-AI Guideline: Overview

- Provides definitions and justifications for six guiding principles
  - **Fairness (1–3)**
  - **Universality (1–4)**
  - **Traceability (1–6)**
  - **Usability (1–5)**
  - **Robustness (1–3)**
  - **Explainability (1–2)**
  - **General (1–7)**
- Presents 30 recommendations for trustworthy AI in healthcare
- Table 2 summarises recommendations and compliance levels
  - + recommended
  - ++ highly recommended
- Supplementary Table 1: glossary of terms
- Supplementary Table 2: relevant stakeholders.



Overall Methodology

Clusters of requirements	Core principles
1. Fairness, diversity, inclusiveness, non-discrimination, unbiased AI, equity	Fairness
2. Generalisability, adaptability, interoperability, applicability, universality	Universality
3. Traceability, monitoring, continuous learning, auditing, accountability	Traceability
4. Human centred AI, user engagement, usability, accessibility, efficiency	Usability
5. Robustness, reliability, resilience, safety, security	Robustness
6. Transparency, explainability, interpretability, understandability	Explainability

Table 1 | Clustering of trustworthy artificial intelligence (AI) requirements and selection of FUTURE-AI guiding principles

Table 2 | List of FUTURE-AI recommendations, together with the expected compliance for both research and deployable artificial intelligence (AI) tools

(+ : recommended ‘++ : highly recommended)

Category	Recommendations	Research	Deployable
Fairness	1. Define any potential sources of bias from an early stage	++	++
	2. Collect information on individuals’ and data attributes	+	+
	3. Evaluate potential biases and, when needed, bias correction measures	+	++
Universality	1. Define intended clinical settings and cross-setting variations	++	++
	2. Use community-defined standards (eg, clinical definitions, technical standards)	+	+
	3. Evaluate using external datasets and/or multiple sites	++	++
	4. Evaluate and demonstrate local clinical validity	+	++
Traceability	1. Implement a risk management process throughout the AI lifecycle	+	++
	2. Provide documentation (eg, technical, clinical)	++	++
	3. Define mechanisms for quality control of the AI inputs and outputs	+	++
	4. Implement a system for periodic auditing and updating	+	++
	5. Implement a logging system for usage recording	+	++
	6. Establish mechanisms for AI governance	+	++
Usability	1. Define intended use and user requirements from an early stage	++	++
	2. Establish mechanisms for human-AI interactions and oversight	+	++
	3. Provide training materials and activities (eg, tutorials, hands-on sessions)	+	++
	4. Evaluate user experience and acceptance with independent end users	+	++
	5. Evaluate clinical utility and safety (eg, effectiveness, harm, cost-benefit)	+	++
Robustness	1. Define sources of data variation from an early stage	++	++
	2. Train with representative real-world data	++	++
	3. Evaluate and optimise robustness against real-world variations	++	++
Explainability	1. Define the need and requirements for explainability with end users	++	++
	2. Evaluate explainability with end users (eg, correctness, impact on users)	+	+
General	1. Engage interdisciplinary stakeholders throughout the AI lifecycle	++	++
	2. Implement measures for data privacy and security	++	++
	3. Implement measures to address identified AI risks	++	++
	4. Define adequate evaluation plan (eg, datasets, metrics, reference methods)	++	++
	5. Identify and comply with applicable AI regulatory requirements	+	++
	6. Investigate and address application-specific ethical issues	+	++
	7. Investigate and address social and societal issues	+	+

Fairness: Definition

- AI tools should maintain similar performance across individuals and groups
- Includes under-represented and disadvantaged groups
- Biases may arise from:
  - individual attributes
  - data acquisition and processing
- Perfect fairness may be impossible
- Biases should be identified, reported, and minimised

Fairness: Recommendations (1–3)

- Fairness 1: Define potential sources of bias from an early stage
- Fairness 2: Collect information on individuals' and data attributes
- Fairness 3: Evaluate potential biases and apply mitigation measures when needed

Universality: Definition

- AI tools should maintain similar performance across individuals and groups
- Includes under-represented and disadvantaged groups
- Biases may arise from:
  - individual attributes
  - data acquisition and processing
- Perfect fairness may be impossible
- Biases should be identified, reported, and minimised

Universality: Recommendations (1–4)

- Universality 1: Define intended clinical settings and cross-setting variations
- Universality 2: Use community defined standards (eg, clinical definitions, technical standards)
- Universality 3: Evaluate using external datasets and/or multiple sites
- Universality 4: Evaluate and demonstrate local clinical validity

Traceability: Definition

- Document and monitor the full AI lifecycle
- From development to deployment and usage
- Increase transparency and accountability
- Enable continuous auditing and updating

Traceability: Recommendations (1–6)

- Traceability 1: Implement a risk management process throughout the AI lifecycle
- Traceability 2: Provide documentation (eg, technical, clinical)
- Traceability 3: Define mechanisms for quality control of the AI inputs and outputs
- Traceability 4: Implement a system for periodic auditing and updating
- Traceability 5: Implement a logging system for usage recording
- Traceability 6: Establish mechanisms for AI governance

Usability: Definition

- End users should use AI efficiently and safely
- Easy interaction with minimal errors
- Clinically useful and safe
- Should improve outcomes and avoid harm

Usability: Recommendations (1–5)

- Usability 1: Define intended use and user requirements from an early stage
- Usability 2: Establish mechanisms for human-AI interactions and oversight
- Usability 3: Provide training materials and activities (eg, tutorials, hands-on sessions)
- Usability 4: Evaluate user experience and acceptance with independent end users
- Usability 5: Evaluate clinical utility and safety (eg, effectiveness, harm, cost-benefit)

Robustness: Definition

- Maintain performance under real-world variations
- Variations may be expected or unexpected
- Small changes can lead to incorrect decisions

Robustness: Recommendations (1–3)

- Robustness 1: Define sources of data variation from an early stage
- Robustness 2: Train with representative real world data
- Robustness 3: Evaluate and optimise robustness against real world variations

Explainability: Definition & Recommendations (1–2)

- **Provide clinically meaningful explanations**
  - Explainability 1: Define the need and requirements for explainability with end users
  - Explainability 2: Evaluate explainability with end users (eg, correctness, impact on users)

General: Definition & Recommendations (1–7)

- General 1: Engage interdisciplinary stakeholders throughout the AI lifecycle
- General 2: Implement measures for data privacy and security
- General 3: Implement measures to address identified AI risks
- General 4: Define adequate evaluation plan (eg, datasets, metrics, reference methods)
- General 5: Identify and comply with applicable AI regulatory requirements
- General 6: Investigate and address application specific ethical issues
- General 7: Investigate and address social and societal issues

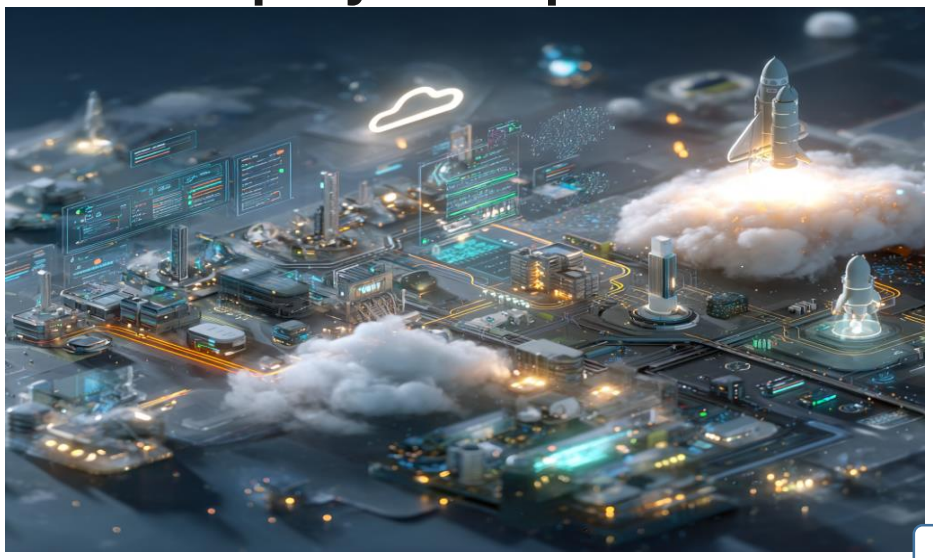

Operationalisation of FUTURE-AI

1- Design phase

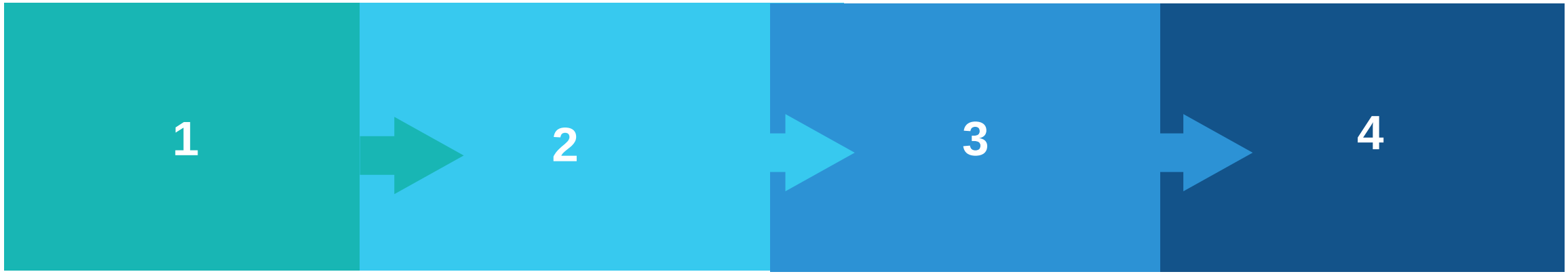
2- Development phase

3- Validation phase

4- Deployment phase



Operationalizing Principles and Recommendations Across Four Phases



- Human-centred design
- Early risk identification
- Stakeholder engagement
- Definition of intended use & risks

- Representative real-world data
- Variability across centers
- Risk mitigation
- Human-in-the-loop mechanisms

- Beyond performance
- Multidimensional trustworthy AI evaluation
- Documentation for transparency & traceability

- Local clinical validation
- User training
- Post-deployment monitoring
- Regulatory alignment


Discussion

Why FUTURE-AI?

- Despite rapid AI advances, real-world clinical adoption remains limited
  - Trust, ethics, safety, and accountability are major barriers
  - Performance alone is insufficient for deployment in healthcare

- FUTURE-AI addresses this gap
  - Focuses on trustworthy and deployable AI
  - Aligns technical development with clinical, ethical, and regulatory needs

- Lifecycle-based framework
  - Covers design → development → validation → deployment & monitoring
  - Goes beyond reporting or post-hoc evaluation guidelines

- Multidimensional trust
  - Technical robustness
  - Clinical safety and utility
  - Ethical, social, and legal considerations

FUTURE-AI as a Living Framework

Not a Static Guideline:

- FUTURE-AI is designed as a living framework
  - To evolve with technology, regulation, and practice

- Continuous refinement
  - Based on user feedback and emerging challenges

Dissemination and Community Engagement:

- Active dissemination strategy
  - Official website: www.future-ai.eu
  - Webinars and community interaction

- Target stakeholders
  - Researchers
  - Clinicians
  - Healthcare organizations
  - Policymakers and regulators

Limitation & Open Challenges

General Guideline, Not Domain-Specific:

- FUTURE-AI is intentionally generic
  - Designed to be applicable across healthcare AI domains

- Requires domain-specific operationalization
  - Radiology, pathology, CDS, genomics need tailored extensions

Regulatory and Legal Uncertainty:

- AI regulations are evolving
  - Example: EU AI Act still under development

- Open issues remain
  - Liability and responsibility
  - Model updates and fine-tuning after deployment
  - Post-deployment accountability

Limitation & Open Challenges

Feasibility, Cost, and Equity Concerns:

- Full compliance may be resource-intensive
  - Especially challenging in low-resource settings
- Trade-off exists
  - Ideal trustworthiness vs real-world feasibility
- Third-party evaluation
  - Desirable in principle
  - Often impractical at scale or globally

Implications for Practice & Policy

For AI Developers:

- Roadmap for trustworthy AI
  - Early risk identification
  - Reduced late-stage redesign and failure

For Clinicians & Healthcare Systems:

- Improved transparency and accountability
  - Clear documentation and governance

- Enhanced trust in AI-supported decisions

For Policymakers & Regulators:

- Bridge between regulation and implementation
  - Supports harmonization of standards
  - Facilitates responsible AI adoption in healthcare


Thank
YOU