UNIVERSITÄT BERN

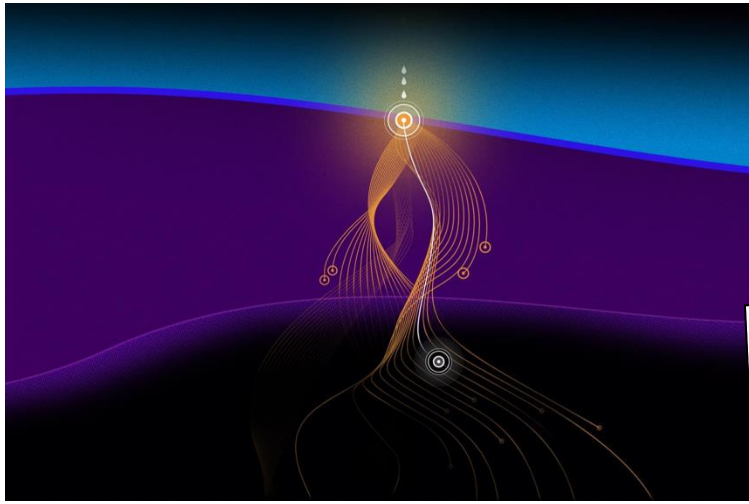
INSELGRUPPE

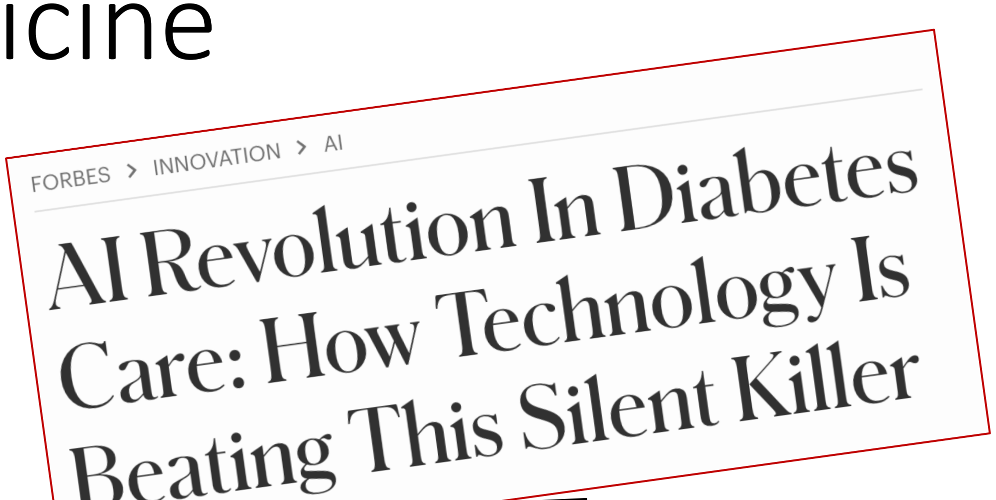
DCB Global Center for Technology Innovation in Diabetes

# Trustworthy AI in Medicine: Introduction

Lisa Koch, 19 February 2025

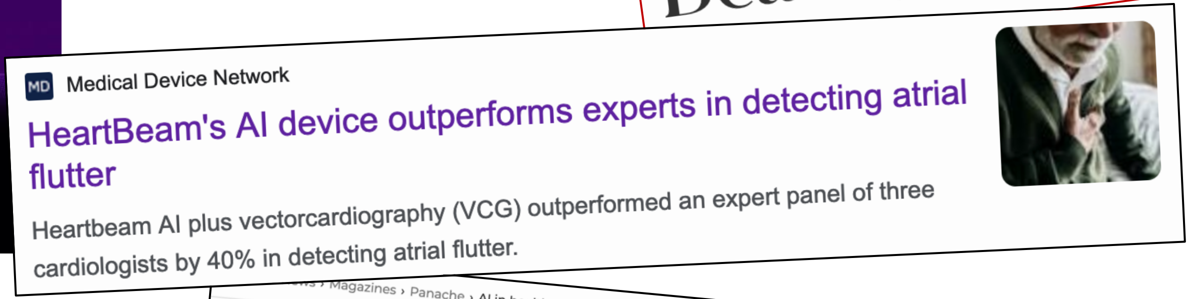
# Artificial Intelligence in Medicine





FORBES › INNOVATION › AI

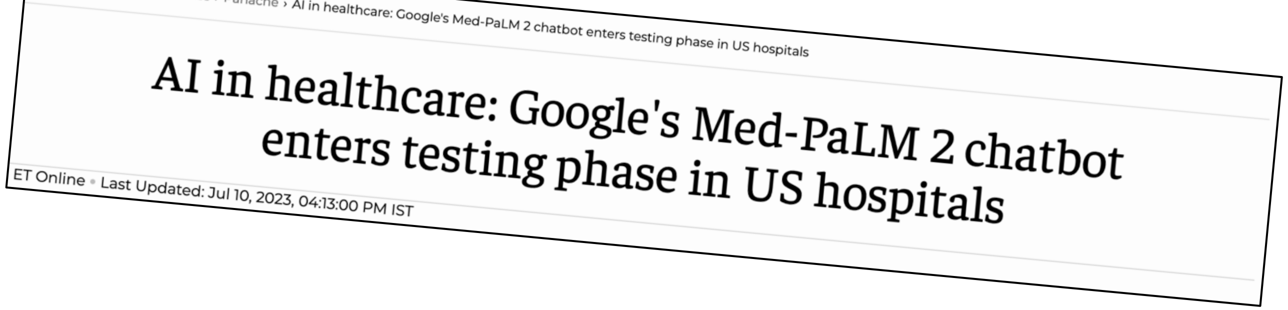
AI Revolution In Diabetes Care: How Technology Is Beating This Silent Killer



MD Medical Device Network

HeartBeam's AI device outperforms experts in detecting atrial flutter

Heartbeam AI plus vectorcardiography (VCG) outperformed an expert panel of three cardiologists by 40% in detecting atrial flutter.

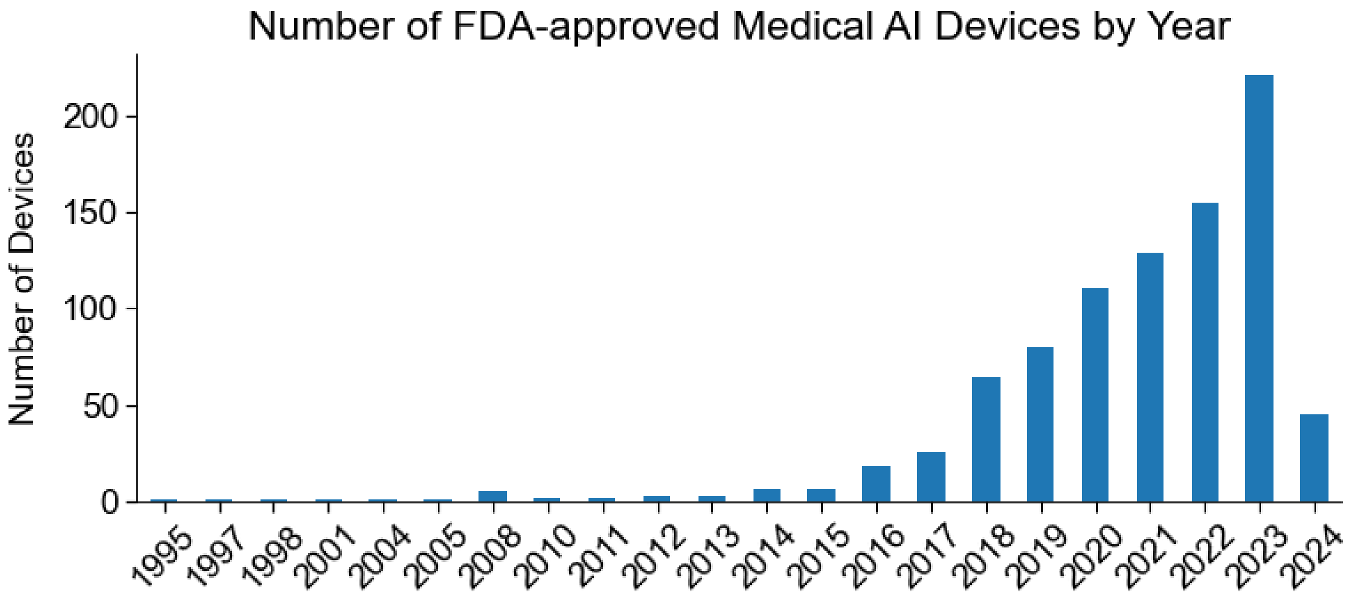


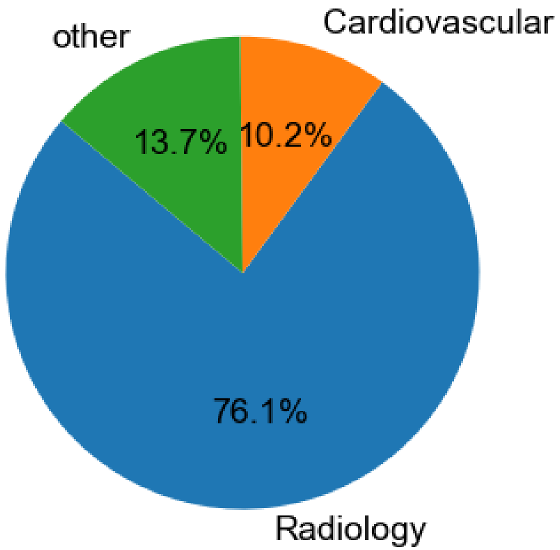
Magazines › Panache › AI in healthcare: Google's Med-PaLM 2 chatbot enters testing phase in US hospitals

AI in healthcare: Google's Med-PaLM 2 chatbot enters testing phase in US hospitals

ET Online • Last Updated: Jul 10, 2023, 04:13:00 PM IST

# AI assisted medical devices approved by FDA



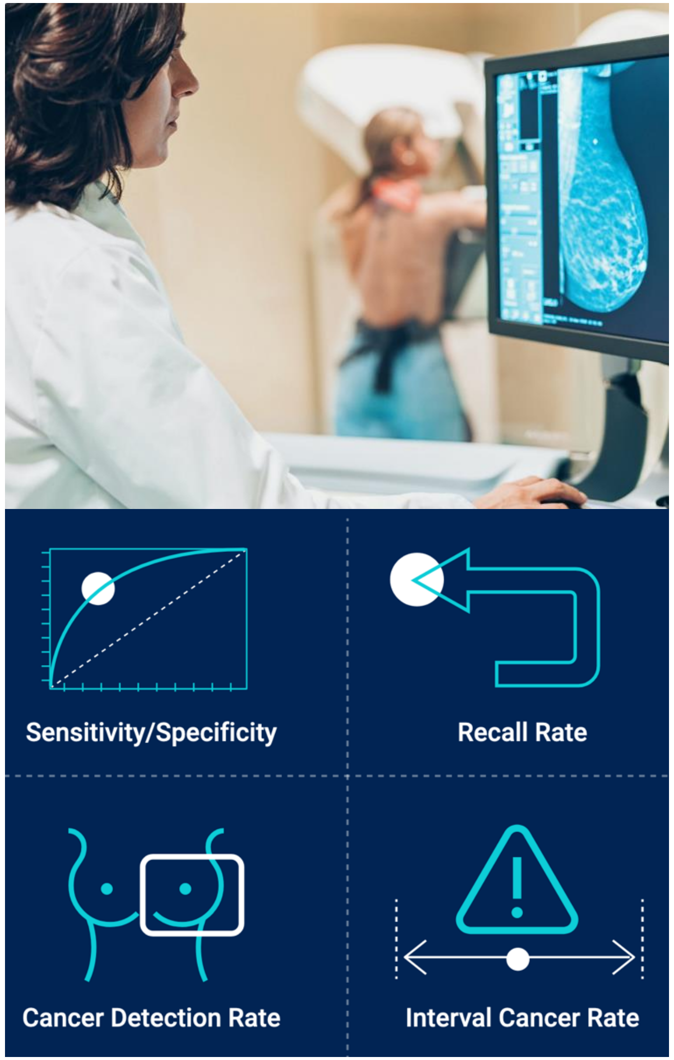


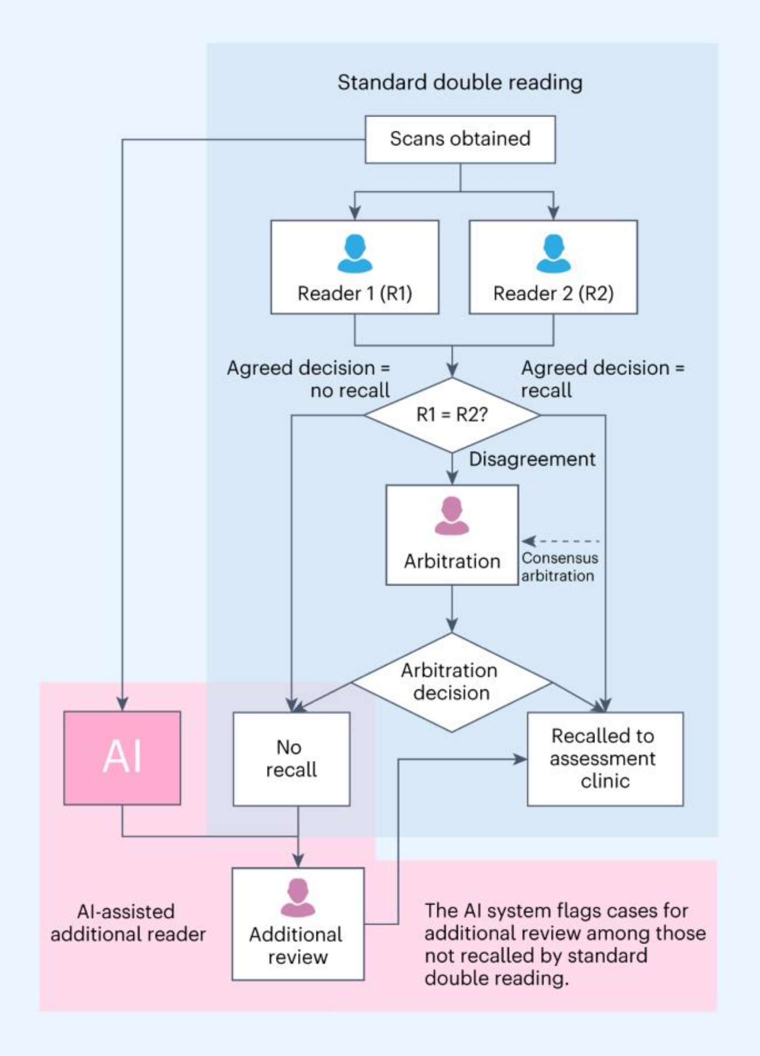
Source: FDA (www.fda.gov), Oct 2024

# Example: breast cancer screening

- AI as additional reader
- Detected 12% more positive cancer cases
- Few additional false positives

- Commercially available
- CE (EU, UK)




Standard double reading

Scans obtained

Reader 1 (R1)

Reader 2 (R2)

Agreed decision = no recall

R1 = R2?

Agreed decision = recall

Disagreement

Arbitration

Consensus arbitration

Arbitration decision

AI

No recall

Recalled to assessment clinic

AI-assisted additional reader

Additional review

The AI system flags cases for additional review among those not recalled by standard double reading.

# Example: Early detection of diabetic retinopathy

LUMINETICSCORE™

AI Diagnostic System *Formerly IDx-DR*

# Potential consequences of AI mistakes

Inefficiency:

- Time and money

**Patient harm:**

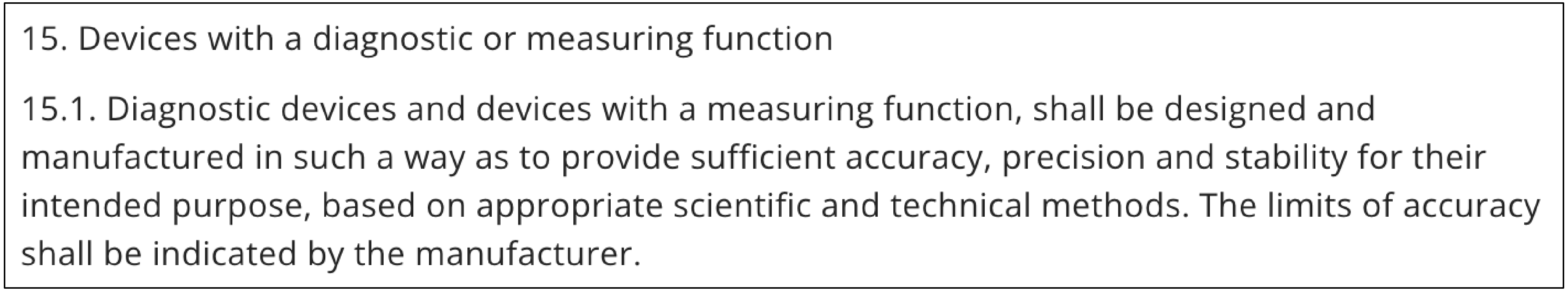
- False negative: inadequate treatment, increased morbidity, death
- False positive: decreased quality of life, personal and public health costs

AI tools for patient care are regulated as medical devices (Software as a Medical Device SaMD)

- Medical Device Regulation (MDR, EU), FDA (USA)
- *Risk class* depends on level of concern, potential harm

# Regulatory frameworks, in a nutshell (EU)

- **Medical device regulation** (MDR, EU)
  - Not very specific!
  - Example from Annex I (General safety and performance requirements):



15. Devices with a diagnostic or measuring function

15.1. Diagnostic devices and devices with a measuring function, shall be designed and manufactured in such a way as to provide sufficient accuracy, precision and stability for their intended purpose, based on appropriate scientific and technical methods. The limits of accuracy shall be indicated by the manufacturer.

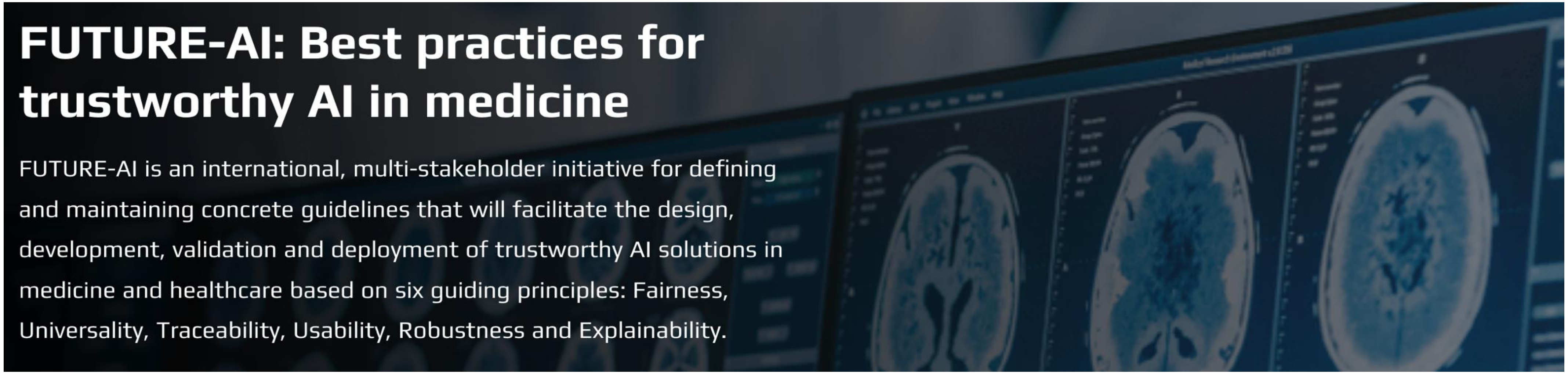
# Regulatory frameworks, in a nutshell (EU)

- **Harmonised standards:**
  - Following standards makes it easier to demonstrate conformity with MDR, safety and performance of device
  - Practical guidance and implementation details to achieve complicance
  - Examples:
    - EN ISO 13485:2016 Quality Management System
    - EN ISO 14971:2019 Risk Management
  - Following standards is not mandatory!

# Gaps in harmonised standards

- Harmonised standards slow to adopt recent developments in AI

- **Consensus guidelines:** preliminary, informal, “pre-standard”



## FUTURE-AI: Best practices for trustworthy AI in medicine

FUTURE-AI is an international, multi-stakeholder initiative for defining and maintaining concrete guidelines that will facilitate the design, development, validation and deployment of trustworthy AI solutions in medicine and healthcare based on six guiding principles: Fairness, Universality, Traceability, Usability, Robustness and Explainability.

Lekadir et al. (2025) *BMJ*

# What about Switzerland?

(Mostly) harmonised with EU regulation:

- MDR (EU) – medical devices regulations (CH)
- GDPR (EU) - nFADP (CH)
- Caution: some alignment gaps exist!

# ML-specific risk management

**Real-world risks when deploying ML models**

- Model performs worse for some patient groups
- prediction is silently wrong
- heterogeneous data quality
- system is used incorrectly
- Sensitive data can be extracted from the model
- real-world data is different from clinical validation data

Risk management

**Mitigations: Trustworthy ML technology**

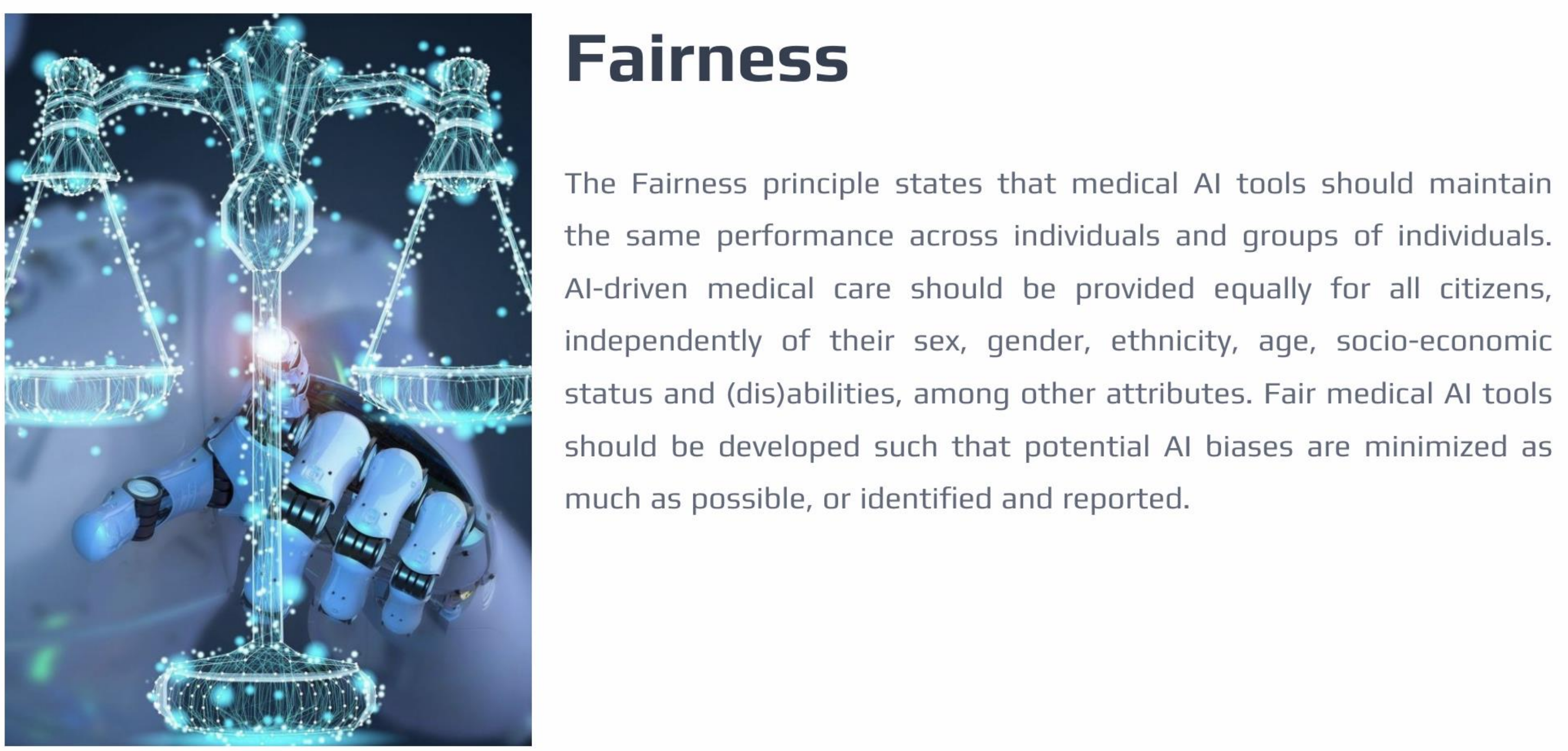
- Fairness
- Reliability, outlier detection
- Robustness
- Explainability
- Privacy and security
- Validation

# Fairness: a possible definition

“A **concept for defining, quantifying and mitigating unfairness** from machine-learning predictions that may cause **disproportionate harm to individuals or groups** of individuals.”

Chen et al. (2023) *Nat Biomed Eng*

# Fairness: according to FUTURE-AI



## Fairness

The Fairness principle states that medical AI tools should maintain the same performance across individuals and groups of individuals. AI-driven medical care should be provided equally for all citizens, independently of their sex, gender, ethnicity, age, socio-economic status and (dis)abilities, among other attributes. Fair medical AI tools should be developed such that potential AI biases are minimized as much as possible, or identified and reported.

From: https://future-ai.eu/principle/fairness/ (17 Feb 2025)

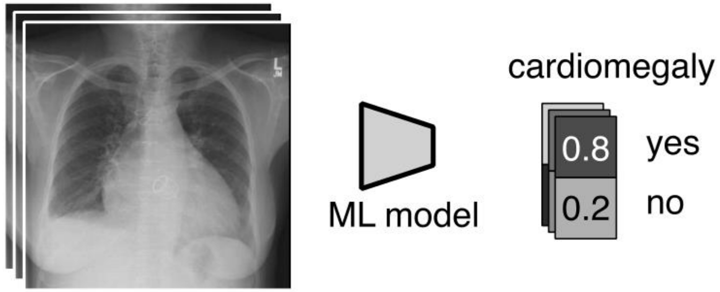
# Who requires fairness?

- **Medical device regulations (EU):** Requirements for „fairness“ ambiguous
- **EU AI act:** Contrast between technical terminology around fairness and law
- **Laws of non-discrimination**
- **Patients, the public, expert groups:** advocacy is required to create standards and inform policy

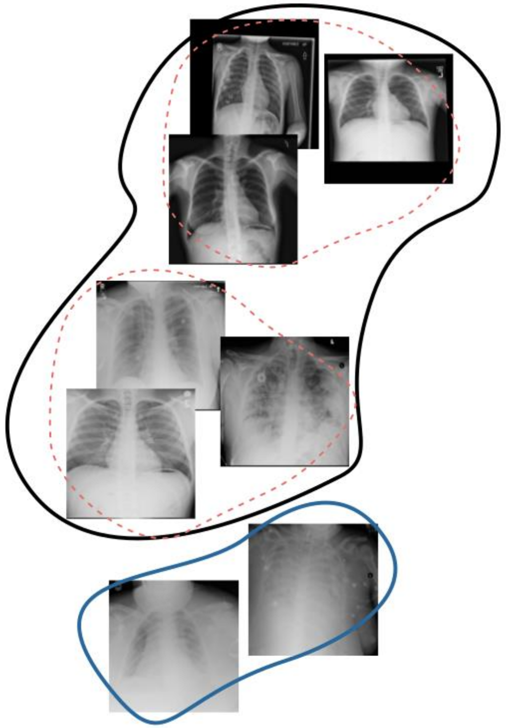
# Fairness: research directions

- Analysis: Expose inequalities in AI-assisted patient care
- Bias mechanisms: What are biases? How do biases arise in healthcare settings?
- Fairness metrics
- Methods for bias detection
- Methods for bias mitigation

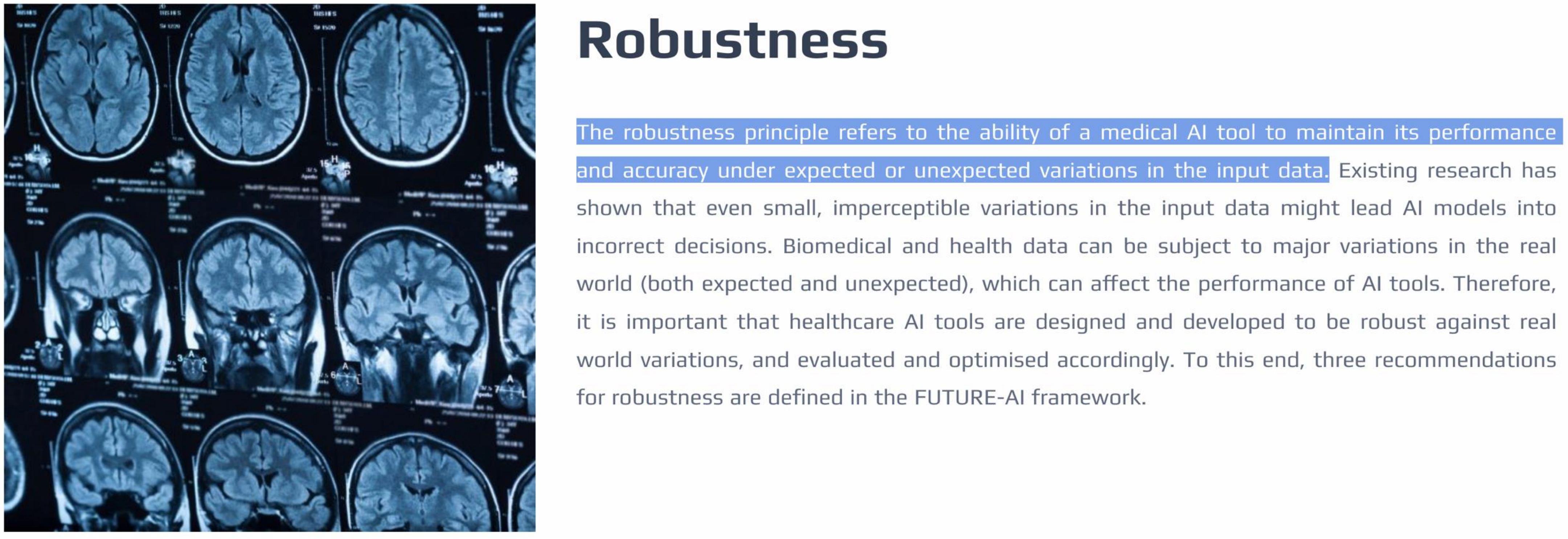
# Reliability



Deep learning models can fail (silently) on data that is not well represented in the training data



- Detect outliers
- Detect failures
- Detect and characterise distribution shifts



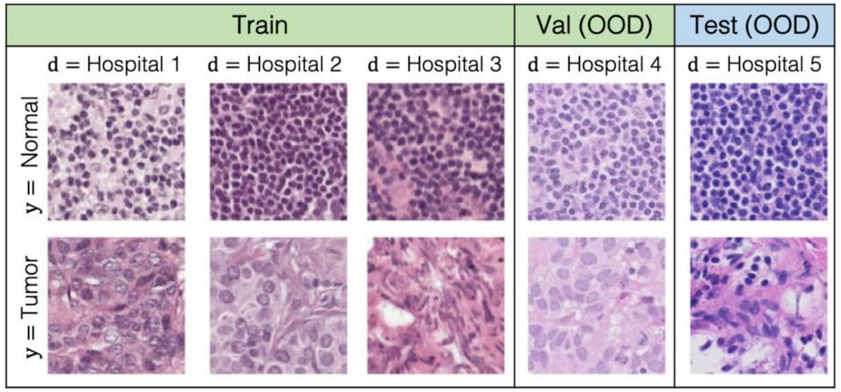
# Robustness

The robustness principle refers to the ability of a medical AI tool to maintain its performance and accuracy under expected or unexpected variations in the input data. Existing research has shown that even small, imperceptible variations in the input data might lead AI models into incorrect decisions. Biomedical and health data can be subject to major variations in the real world (both expected and unexpected), which can affect the performance of AI tools. Therefore, it is important that healthcare AI tools are designed and developed to be robust against real world variations, and evaluated and optimised accordingly. To this end, three recommendations for robustness are defined in the FUTURE-AI framework.

From: https://future-ai.eu/principle/fairness/ (18 Feb 2025)

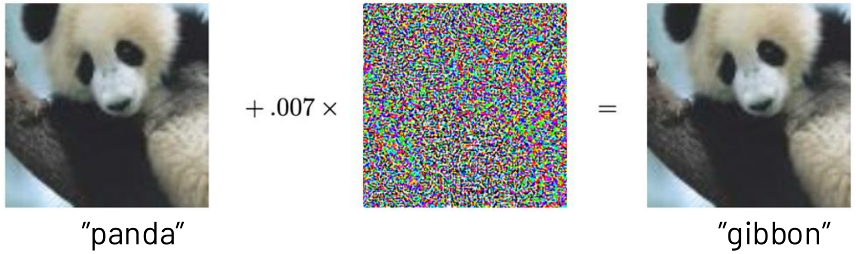
# (Some) dimensions of robustness

## Domain generalisation



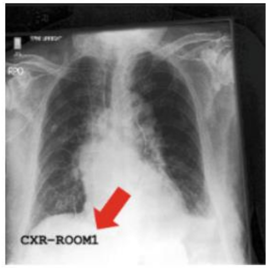
Koh et al. (2021) *Proc. ICML*

## Adversarial robustness



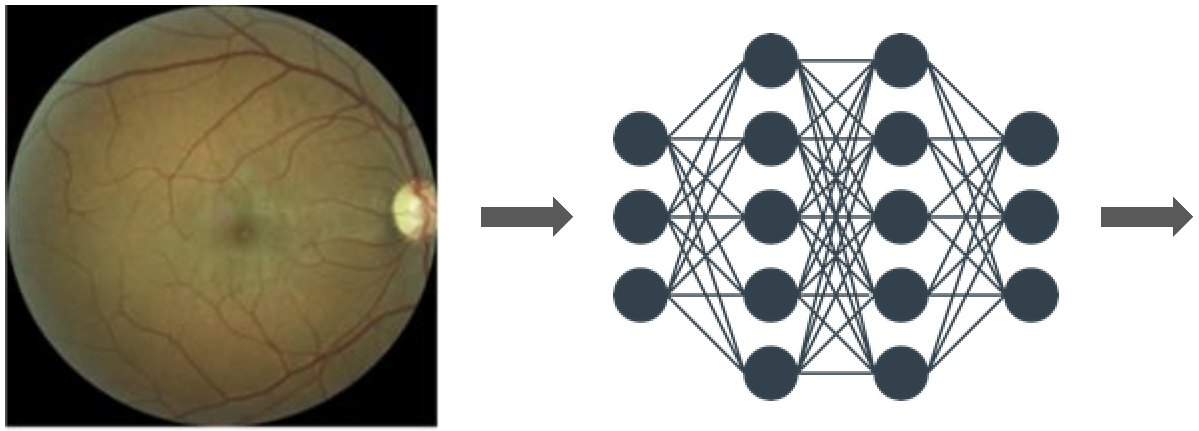
Goodfellow et al. (2015) *Proc. ICLR*

## Shortcut learning



Sun et al. (2023) *Proc. MICCAI*

# Explainability



diseased

~~healthy~~

**Transparency in high stakes applications:**

- trust affects downstream decisions
- inspection and certification
- mandated by GDPR, recommended in guidance documents

# Explainability: regulations

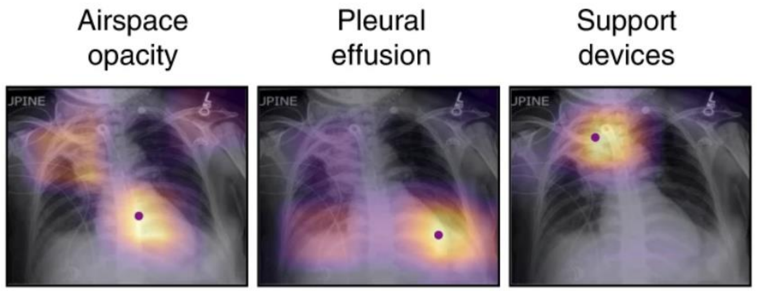
- **EU General Data Protection Regulation:** „[the data subject should have] the right … to obtain an explanation of the decision reached“

- **EU AI act (Art. 171):** „Affected persons should have the right to obtain an explanation where a deployer’s decision is based mainly upon the output from certain high-risk AI systems …“

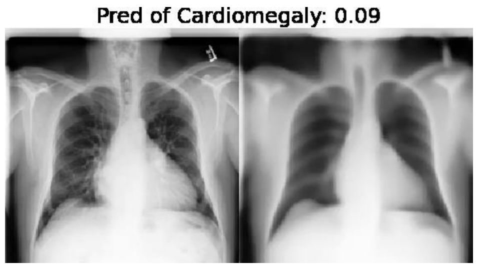
- **MDR (Annex II, 1.1):** “principles of operation“, "a general description of the key functional elements, e.g. its parts/components (including software if appropriate), its formulation, its composition, its functionality…“

# Approaches to explainable AI

- Post-hoc:
    - **Feature attribution**
      images: saliency maps)
    - **Counterfactual explanations**
      What if… it were diseased? Healthy?
- Inherently interpretable models: explain mechanism directly



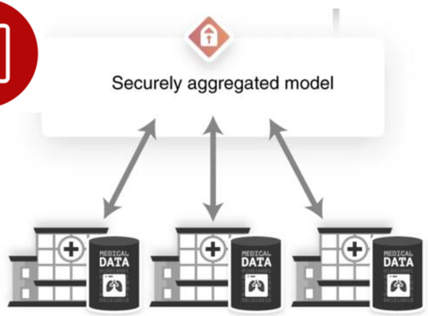
Saporta et al. (2022) *Nature Machine Intelligence*



Cohen et al. (2021) *Proc. MIDL*

# Privacy and security

- Concerns during **development phase**:
  - data protection of patients providing training data (e.g. *membership inference*) → Privacy-preserving deep learning, differential privacy
  - commercial interests and hospital guidelines: private datasets → Federated learning: data never leaves the hospital



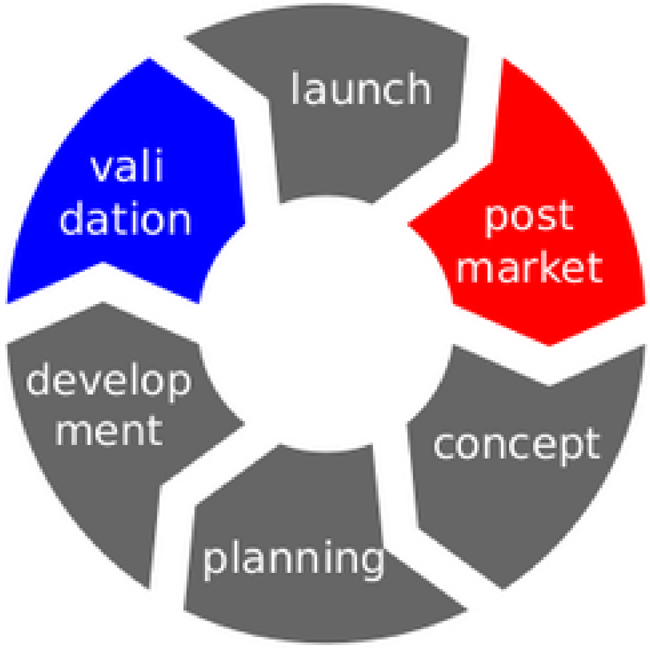
Kaissis et al. (2022) *NatMachInt*

- Concerns during **deployment phase**:
  - Data protection of patients
  - Security: prevent tampering with model predictions

→ Cybersecurity
Encrypted inference

# Validation

- Thorough preclinical and clinical evaluation
- Reflect intended use: data and application scenario
- Subgroup analysis, systematic error exploration
- Relevant performance metrics

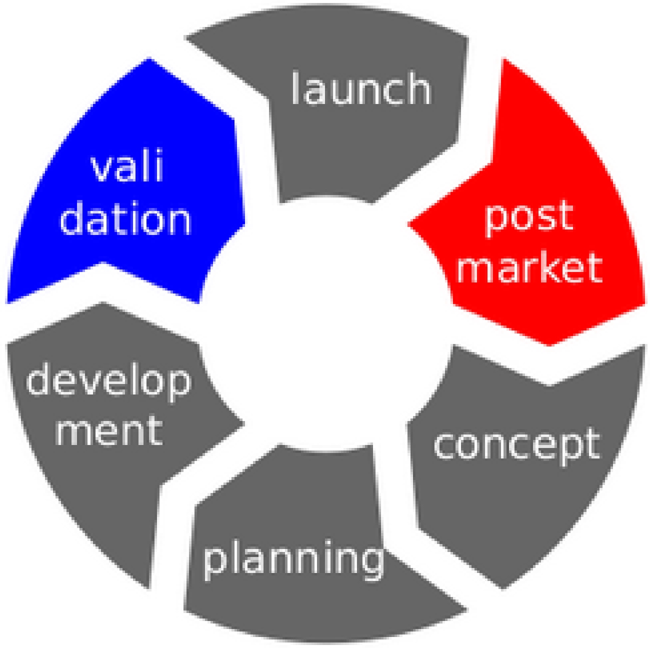


# Clinical validation is not enough

Performance claims from clinical validation studies are not enough to answer:

**How well will a ML model work for *during deployment?***

Increased emphasis on postmarket surveillance



# Paper selection in this seminar

- Cover wide range of topics in trustworthy AI
- Includes some **non-medical seminal work** with high impact on AI in medicine. Put these into context.