ESC European Heart Journal (2026) **00**, 1–2
https://doi.org/10.1093/eurheartj/ehag303

CardioPulse

Global Spotlights

Trustworthiness of artificial intelligence from a patient perspective

Axel Verstrael [1,2,*], **Jennifer Camaradou** [2], and **Bart Scheenaerts** [2]

[1]Faculty of Medicine and Life Sciences, Research Group, Healthcare & Ethics, University of Hasselt, Agoralaan Gebouw D, Diepenbeek B-3590, Belgium; and [2]ESC Patients Forum, European Society of Cardiology, Les Templiers 2035 route des Colles CS 80179 Biot 06903, Sophia Antipolis Cedex, France

Discussions about the trustworthiness of artificial intelligence (AI) in healthcare are increasingly grounded in ethical frameworks rather than technical performance metrics.

A key reference is Ethics and governance of artificial intelligence for health [World Health Organization (WHO)],[1] which formulates six ethical consensus principles for: protecting human autonomy; promoting well-being, safety, and the public interest; ensuring transparency, explainability, and intelligibility; fostering responsibility and accountability; ensuring inclusiveness and equity; and promoting responsive and sustainable AI. These principles do not primarily describe how healthcare systems should be organized, but rather define normative design, implementation, and governance principles.

From the patient's perspective, these principles resonate fundamentally with the inherent nature of medical practice: becoming a patient already means coming into a system structured around uncertainty, probability, and imperfect knowledge. Artificial intelligence can introduce and remove uncertainty in healthcare[2] paradoxically by entering an intrinsically uncertain domain while simultaneously supporting risk–benefit calculations with predictable modelling.

This management of uncertainty is crucial because medical decisions are rarely absolute; diagnoses, prognoses, and treatment outcomes are inherently probabilistic. For patients, uncertainty is not a theoretical construct but a daily reality. Making this uncertainty explicit, for instance, through AI systems displaying confidence intervals, reflects a form of epistemic honesty. Such transparency can strengthen trust by aligning technological outputs with the patient's lived experience of medical ambiguity, thereby enhancing shared decision-making.

The clinical necessity of addressing this ambiguity is underscored by the high stakes of medical error: model-based extrapolations from WHO data suggest medication errors may be associated with roughly 163 000 deaths annually in the EU, while digitally integrated medication-traceability systems show error reduction up to 58%.[3,4] Such figures illustrate that error and system failure are characteristics of complex healthcare environments. From the patient's viewpoint, trust tends to be placed in imperfect systems. Therefore, the ethical challenge is not about AI's flawlessness, but about a design that makes uncertainties visible, accountable, and manageable.

Balancing human and machine fallibility

This alignment between design and reality extends to the principle of autonomy, where the fallibility of both human and machine must be balanced: autonomy in healthcare is not merely about informed consent, but about preserving human agency within technological systems. Requiring clinicians to be able to override AI without friction recognizes human vulnerability and contextual judgement. Patients understand clinicians are subject to fatigue and cognitive overload. After performing extensive surgery, physicians may experience mental exhaustion or reduced attention, states that can lead to clinical lapses analogous to the 'hallucinations' found in AI models. In these high-pressure contexts, AI support functions as a vital cognitive aid rather than a replacement, mitigating the risks of human fallibility and enhancing patient safety.

In critical moments, this support is often prioritized over technical perfection, as the WHO's focus on safety reframes the idea of technological perfection. Whilst facing severe illness, patients do not have the privilege to wait for ideal conditions; they welcome any system capable of providing rapid screening and therapeutic possibilities, even if imperfect. Hence, AI can be experienced as an epistemic amplification, which aids timely decision-making under pressure.

This amplification is particularly valuable for complex cases, as AI may offer particular benefit in patients living with complex

* Corresponding author. Email: axel.verstrael@uhasselt.be

multimorbidity facing fragmented care pathways. Integrative AI systems capable of synthesizing longitudinal, multimodal data may support coherent, personalized management strategies.[5]

However, the successful integration of these systems depends heavily on how they mitigate broader systemic risks, as research discussed by Camaradou[6] shows that patients generally support AI where it enables faster diagnosis and improved outcomes. However, support depends less on technical properties like reproducibility metrics than on potential harm mitigation. Patients seek an understandable explanation of risks and reassurance that clinicians remain responsible. Yet the ability of clinicians to exercise this responsibility is challenged by emerging evidence regarding automation bias and the potential deskilling effect of prolonged AI exposure.[7]

Addressing structural bias and inherited inequities

Beyond individual clinical risks, harm mitigation must also address deep-seated structural biases: the WHO's principle of inclusiveness and equity becomes pivotal when considering bias. Pharmaceutical research has long been underrepresented by women, with a lack of diversity overall, as randomized clinical trials frequently focus on adult men, with findings being generalized across sex, age, and body composition. This remains a pressing issue: recent cardiovascular research demonstrates that women tend to experience gender-specific risks, including inappropriate dosing, delayed treatment, and bias in clinical decision-making.[8] Artificial intelligence does not create this problem but inherits it.

The path to trust: transparency and collaborative design

To effectively break this cycle of inherited bias, rigorous methodological standards are essential: mitigating concerns regarding data representation and health inequalities requires the application of reporting standards like TRIPOD + AI[9] and CONSORT-AI/SPIRIT-AI,[10] which seek to improve transparency and provide governance. The ethical opportunity lies in making bias visible and debatable. If data sources and model assumptions are made explicit through these standards, they may offer tools to expose and correct long-standing inequities rather than reproduce them.[1]

Ultimately, this transparency serves as the link between the system and its users, as it reinforces the idea that AI systems are used by imperfect users.[11] Much more than a technical requirement, transparency is a foundation for trust, determining the distribution of responsibilities and negotiation of uncertainty within therapeutic care relationships.

Within this ethical landscape, a shift towards collaborative design is necessary: concepts like human-centred AI[12] and human-in-the-loop become the norm rather than design refinements. They implement the WHO's principles by insisting AI supports human judgement, responds to human needs, and upholds human values. They also open the space for patients during the design of AI, providing end-users with feedback and co-design to shape systems for care trajectories.[5]

In conclusion, for patients, trustworthy AI is not therefore defined by technical perfection, but by an ethical alignment with health-related reality: consisting of uncertainty, imperfection, complexity, and the continuous negotiation of trust. Artificial intelligence in healthcare does not provide certainty; it forms part of an ambiguous world. Ethical design acknowledges this inaccuracy and manages it responsibly, ensuring technology remains oriented towards human well-being. The ultimate goal is to ensure that healthcare remains a deeply human experience.

Declarations

Disclosure of Interest

Nothing to declare.

References

1. World Health Organization. *Ethics and governance of artificial intelligence for health: WHO guidance*. Geneva: World Health Organization; 2021. Licence: CC BY-NC-SA 3.0 IGO.
2. Tauben Averbuch A, Asselbergs FW, Vardas P, Van Spall HGC. Great debate: artificial intelligence will replace much of what cardiologists do. *Eur Heart J* 2025;**46**:3628–3635. https://doi.org/10.1093/eurheartj/ehaf305
3. European Alliance for Access to Safe Medicines (EAASM). Medication errors - the most common adverse event in hospitals threatens patient safety and causes 160,000 deaths per year. EAASM Press Release; 2022 Sep 13 [cited 2026 Apr 10]. Available from: https://eaasm.eu/en-gb/2022/09/13/press-release-medication-errors-the-most-common-adverse-event-in-hospitals-threatens-patient-safety-and-causes-160000-deaths-per-year/
4. World Health Organization. *Global patient safety action plan 2021–2030: towards eliminating avoidable harm in health care*. Geneva: World Health Organization; 2021. Licence: CC BY-NC-SA 3.0 IGO.
5. Biswas D. Artificial intelligence for cardiovascular care in action: from learning to implementation in health systems. *JACC Adv* 2025;**4**:101740. https://doi.org/10.1016/j.jacadv.2025.102307
6. Camaradou J. Patient perspectives on the "Black Box": transparency and accountability in AI diagnostics. *Patient Exp J* 2023;**40**:2563–2572. https://doi.org/10.1007/s12325-023-02511-3
7. Budzyń K, Romańczyk M, Kitala D, Kołodziej P, Bugajski M, Adami HO, *et al.* Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicenter, observational study. *Lancet Gastroenterol Hepatol* 2025; **10**:e12. https://doi.org/10.1016/S2468-1253(25)00294-8
8. Paradies V, Masiero G, Rubboli A, Van Beusekom HMM, Costa F, Capranzano P, *et al.* Antithrombotic drugs for acute coronary syndromes in women: sex-adjusted treatment and female representation in randomised clinical trials. *EuroIntervention* 2025;**21**:e655–67. https://doi.org/10.4244/EIJ-D-24-00876
9. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, *et al.* TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 2024;**385**: e078378. https://doi.org/10.1136/bmj-2023-078378
10. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, Chan A-W, *et al.* Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med* 2020;**26**:1364–74. https://doi.org/10.1038/s41591-020-1034-x
11. Kostick-Quenet KM, Gerke S. AI in the hands of imperfect users. *NPJ Digit Med* 2022;**5**:197. https://doi.org/10.1038/s41746-022-00737-z
12. Abu Hussein NS, Del Pillar Arias M, Marshall DC, Mullins C, Mwavu R, Oliver S, *et al.* Beyond models: a paradigm shift toward human-centered AI system design. *Crit Care* 2025;**29**:514. https://doi.org/10.1186/s13054-025-05714-y