News and Perspectives

The Right to Understand in Health Care AI

Anshu Ankolekar, JMIR Correspondent

Key Takeaways

- The European Union Artificial Intelligence (AI) Act and General Data Protection Regulation (GDPR) give patients legal grounds to seek explanations of AI-driven medical recommendations, but neither framework specifies what a meaningful explanation requires in clinical practice.
- Technical, clinical, and literacy barriers mean that even well-intentioned explanations may not reach patients in a form they can use to make informed decisions about their care.
- Closing this gap requires shifting from compliance (“Was an explanation provided?”) to effectiveness (“Can patients use it?”).

An artificial intelligence (AI) system flags a nodule on a lung computed tomography scan and assigns it an 87% malignancy probability. The radiologist receives a confidence score and a heat map showing which parts of the image the algorithm focused on. When her patient asks, “why does the computer think it’s cancer,” she realizes the output tells her what the AI concluded but not why or how to explain it to her patient.

It can be incredibly challenging to explain AI reasoning in clinical practice. The European Union (EU) AI Act, which entered into force in August 2024 with obligations phasing in through 2027, requires that deployers of high-risk AI systems provide those affected with clear and meaningful explanations of decisions shaped by these systems [1]. This creates a legal basis for patients to seek explanations of AI-driven medical recommendations. It also immediately raises a question the law alone cannot answer: What does a meaningful explanation of an AI medical decision look like in clinical practice, and how do we deliver it?

Legal Foundation and Open Questions

Many clinically deployed medical AI systems will fall into the AI Act’s high-risk category, particularly where they are regulated as medical device software or serve as safety components of regulated products [1,2]. For cases where the AI Act does not directly apply, patients may also draw on the General Data Protection Regulation (GDPR). Under Article 22, the GDPR provides safeguards against decisions “based solely on automated processing,” including the right to meaningful information about the logic behind those decisions [3].

However, most clinical AI does not meet Article 22’s threshold, since a human clinician typically remains the formal decision maker. This creates ambiguity and a paradox: the human oversight meant to protect patients may also reduce their legal claim to explanation, since the decision may no longer be considered purely automated.

Together, these frameworks establish that patients have legal grounds to seek explanations of AI-driven medical recommendations, but these are stronger in principle than in practice, and neither framework specifies what a “meaningful” explanation requires.

Technical Complexity

The difficulty the radiologist in the opening scenario faces is not incidental. The most accurate AI models generate outputs through millions of interacting parameters in ways that even their developers cannot fully trace [4]. Requiring greater transparency can push developers toward simpler, more interpretable models that sacrifice diagnostic accuracy, a trade-off with direct consequences for patient care [5,6].

Current explainable AI methods only partially address this. Saliency maps show which regions of an image the algorithm weighted most heavily, but not what it understood about them [7]. Feature importance rankings indicate which variables mattered, but not how they interacted or why particular thresholds were significant [8]. These post hoc approximations attempt to reconstruct reasoning after the fact [9,10], and they can produce plausible-sounding explanations that do not accurately reflect the model’s internal logic [11].

Clinical Implementation Challenges

Even where technical explanations exist, delivering them in clinical practice has its own barriers. Clinicians typically receive AI outputs as confidence scores and recommendations rather than reasoning, meaning they may be asked to explain a decision they do not fully understand themselves [12,13]. Clinicians already struggle to find time for thorough clinical conversations, and AI adds another layer of complexity to encounters that are already stretched [14-16].

Automation bias—deferring to algorithmic recommendations even when these conflict with clinical judgment—is another implementation challenge clinicians face [17]. A prospective study of radiologists reading mammograms found that incorrect AI suggestions pulled readers toward an incorrect diagnosis regardless of level of experience [18]. An explanation delivered by a clinician who has already deferred to the algorithm may reflect the AI's conclusion rather than an independent clinical assessment, challenging the assumption in both legal frameworks that human oversight guarantees meaningful review.

Patient Understanding

Even if clinicians provide explanations, understanding is not guaranteed. Between 22% and 58% of EU citizens report difficulty accessing, understanding, appraising, and applying the health information they need to navigate health care services, with pronounced gaps among older adults, lower socioeconomic groups, and rural communities [19]. Interpreting AI outputs requires statistical and technical literacy that even a high general education does not guarantee. Many highly educated individuals struggle with medical statistics and probability statements, meaning technically accurate explanations may create barriers regardless of educational background [20].

Providing more technical detail is not likely to help. Research on medical decision-making suggests that excessive technical information can lead to cognitive overload, causing patients to defer to physician authority rather than engage with the explanation [21]. To participate meaningfully in decisions, what patients typically need is not a description of how an algorithm works, but clarity on what's most relevant for their own situation [22,23].

What Implementation Requires

Resolving these challenges requires multiple actors.

Developers could design explanation systems with patient input from the outset, testing comprehension with actual patients rather than demonstrating compliance with legal standards alone [6]. Drawing on principles from shared decision-making [24], risk communication [25], and algorithmic fairness research [26], a useful patient-facing explanation could include what the system is recommending and for what decision point; how confident it is and what that confidence means in practical terms; relevant key limitations, such as known performance gaps in specific populations; and viable alternative options. This reframes the goal from technical transparency to decision-relevant clarity, with effectiveness measurable by whether patients can answer these questions after an encounter.

Health care institutions can also bridge this gap. Allocating time for AI discussions, training staff to support patients in navigating AI-driven recommendations, and establishing clear protocols could shift explanation from a compliance exercise toward genuine patient understanding. Policy makers could support this by developing standards focused on comprehension and investing in digital health literacy programs [2].

Involving patients in the design of explanation systems from the start would strengthen all of these efforts. Patient advocates have highlighted that explanation approaches tend to reflect what developers and regulators consider important, which may not always align with what patients need to know [27]. Co-design partnerships between developers, institutions, and patient communities offer a route toward explanations that are not only legally sound but genuinely useful.

Moving Forward

The EU AI Act gives patients something they did not previously have: a legal basis for demanding transparency about AI systems influencing their care. But the right to explanation and the capacity to deliver one that patients can genuinely use are shaped by forces the law alone cannot govern: the opacity of high-performing models, the pressures of clinical practice, and the diversity of patient needs and literacy levels.

What the AI Act's transparency requirements provide, beyond legal protection, is a shared standard to work toward. The right to explanation is an important starting point. What patients need now are answers they can use.

Keywords: artificial intelligence; government regulation; European Union; clinical decision support systems; informed consent; ethics; patient participation; patient rights; health knowledge, attitudes, practice; health literacy

Conflicts of Interest

None declared.

References

1. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). EUR-Lex. 2024. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689 [Accessed 2026-03-17]
2. van Kolfschooten H, van Oirschot J. The EU Artificial Intelligence Act (2024): implications for healthcare. Health Policy. Nov 2024;149:105152. [doi: 10.1016/j.healthpol.2024.105152] [Medline: 39244818]
3. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive

95/46/EC (General Data Protection Regulation) (Text with EEA relevance). EUR-Lex. 2016. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679 [Accessed 2026-03-17]

4. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. May 2019;1(5):206-215. [doi: 10.1038/s42256-019-0048-x] [Medline: 35603010]
5. Ennab M, Mcheick H. Designing an interpretability-based model to explain the artificial intelligence algorithms in healthcare. Diagnostics (Basel). Jun 26, 2022;12(7):1557. [doi: 10.3390/diagnostics12071557] [Medline: 35885463]
6. Ebers M. AI robotics in healthcare between the EU Medical Device Regulation and the Artificial Intelligence Act: gaps and inconsistencies in the protection of patients and care recipients. Oslo Law Rev. Oct 31, 2024;11(1):1-12. [doi: 10.18261/olr.11.1.2]
7. Borji A. Saliency prediction in the deep learning era: successes and limitations. IEEE Trans Pattern Anal Mach Intell. Feb 2021;43(2):679-700. [doi: 10.1109/TPAMI.2019.2935715] [Medline: 31425064]
8. Hossain MI, Zamzmi G, Mouton PR, Salekin MS, Sun Y, Goldgof D. Explainable AI for medical data: current methods, limitations, and future directions. ACM Comput Surv. Jun 30, 2025;57(6):1-46. [doi: 10.1145/3637487]
9. Bordt S, Finck M, Raidl E, von Luxburg U. Post-hoc explanations fail to achieve their purpose in adversarial contexts. Presented at: FAccT ’22: 2022 ACM Conference on Fairness, Accountability, and Transparency; Jun 21-24, 2022:891-905; Seoul, Republic of Korea. [doi: 10.1145/3531146.3533153]
10. Mhasawade V, Rahman S, Haskell-Craig Z, Chunara R. Understanding disparities in post hoc machine learning explanation. Presented at: FAccT ’24: The 2024 ACM Conference on Fairness, Accountability, and Transparency; Jun 3-6, 2024:2374-2388; Rio de Janeiro, Brazil. [doi: 10.1145/3630106.3659043]
11. Jin Q, Chen F, Zhou Y, et al. Hidden flaws behind expert-level accuracy of multimodal GPT-4 vision in medicine. NPJ Digit Med. Jul 23, 2024;7(1):190. [doi: 10.1038/s41746-024-01185-7] [Medline: 39043988]
12. Sivaraman V, Bukowski LA, Levin J, Kahn JM, Perer A. Ignore, trust, or negotiate: understanding clinician acceptance of AI-based treatment recommendations in health care. Presented at: CHI ’23: CHI Conference on Human Factors in Computing Systems; Apr 23-28, 2023:1-18; Hamburg, Germany. [doi: 10.1145/3544548.3581075]
13. Scott IA, Carter SM, Coiera E. Exploring stakeholder attitudes towards AI in clinical practice. BMJ Health Care Inform. Dec 2021;28(1):e100450. [doi: 10.1136/bmjhci-2021-100450] [Medline: 34887331]
14. Braddock CH, Snyder L. The doctor will see you shortly. The ethical significance of time for the patient-physician relationship. J Gen Intern Med. Nov 2005;20(11):1057-1062. [doi: 10.1111/j.1525-1497.2005.00217.x] [Medline: 16307634]
15. Jacobs M, He J, Pradier MF, et al. Designing AI for trust and collaboration in time-constrained medical decisions: a sociotechnical lens. Presented at: CHI ’21: CHI Conference on Human Factors in Computing Systems; May 8-13, 2021:1-14; Yokohama, Japan. [doi: 10.1145/3411764.3445385]
16. Covvey JR, Kamal KM, Gorse EE, et al. Barriers and facilitators to shared decision-making in oncology: a systematic review of the literature. Support Care Cancer. May 2019;27(5):1613-1637. [doi: 10.1007/s00520-019-04675-7] [Medline: 30737578]
17. Abdelwanis M, Alarafati HK, Tammam MMS, Simsekler MCE. Exploring the risks of automation bias in healthcare artificial intelligence applications: a Bowtie analysis. J Safety Sci Resilience. Dec 2024;5(4):460-469. [doi: 10.1016/j.jnlssr.2024.06.001]
18. Dratsch T, Chen X, Rezazade Mehrizi M, et al. Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance. Radiology. May 2023;307(4):e222176. [doi: 10.1148/radiol.222176] [Medline: 37129490]
19. Collado D. Digital health literacy: a cornerstone of health equity in the EU: policy brief. Health Action International. Oct 2024. URL: https://haiweb.org/wp-content/uploads/2024/10/Digital-Health-Literacy-in-the-EU.pdf [Accessed 2026-02-18]
20. Reyna VF, Nelson WL, Han PK, Dieckmann NF. How numeracy influences risk comprehension and medical decision making. Psychol Bull. Nov 2009;135(6):943-973. [doi: 10.1037/a0017327] [Medline: 19883143]
21. Peters E. Beyond comprehension: the role of numeracy in judgments and decisions. Curr Dir Psychol Sci. 2012;21(1):31-35. [doi: 10.1177/0963721411429960]
22. Hildt E. What is the role of explainability in medical artificial intelligence? A case-based approach. Bioengineering (Basel). Apr 2, 2025;12(4):375. [doi: 10.3390/bioengineering12040375] [Medline: 40281735]
23. Pierce RL, Van Biesen W, Van Cauwenberge D, Decruyenaere J, Sterckx S. Explainability in medicine in an era of AI-based clinical decision support systems. Front Genet. 2022;13:903600. [doi: 10.3389/fgene.2022.903600] [Medline: 36199569]
24. Elwyn G, Frosch D, Thomson R, et al. Shared decision making: a model for clinical practice. J Gen Intern Med. Oct 2012;27(10):1361-1367. [doi: 10.1007/s11606-012-2077-6] [Medline: 22618581]

25. Gigerenzer G, Edwards A. Simple tools for understanding risks: from innumeracy to insight. BMJ. Sep 27, 2003;327(7417):741-744. [doi: 10.1136/bmj.327.7417.741] [Medline: 14512488]

26. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. Oct 25, 2019;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]

27. Kolfschooten HV. EU regulation of artificial intelligence: challenges for patients' rights. Common Market Law Rev. Feb 1, 2022;59(1):81-112. [doi: 10.54648/COLA2022005]

Please cite as:
Ankolekar A
The Right to Understand in Health Care AI
J Med Internet Res 2026;28:e95090
URL: https://www.jmir.org/2026/1/e95090
doi: 10.2196/95090