NEJM
AI

NEJM AI 2026;3(2)
DOI: 10.1056/AIp2501266

PERSPECTIVE

The Missing Dimension in Clinical AI: Making Hidden Values Visible

C. Goldberg,[1,2] R. D. Balicer,[3,4,5] M. Bhat,[6,7] D. Blumenthal,[2,8] R. W. Brendel,[9] E. Brondolo,[10] J. S. Brownstein,[9,11] T. A. Buckley,[9] C. H. Cain,[12] P. Chandak,[13] F. Chessa,[14,15] A. Chopra,[16] N. Dagan,[2,3,4,17] K. S. Ehlert,[18] B. J. Evans,[19] R. Freeman,[20] B. S. Glicksberg,[21] W. Gordon,[9,22] M. I. Gröschel,[23] S. Hoffman,[9] E. M. Hundert,[9] S. Hyare,[24] S. Johri,[9] J. Joseph,[25] M. Klote,[26] A. B. Landman,[27] V. S. Lee,[28] J. C. Mandel,[9,29] K. D. Mandl,[9,11] A. K. Manrai,[2,6] M. Might,[30] G. N. Nadkarni,[21,31] D. J. Nigrin,[14] A. Noori,[9,32] G. S. Omenn,[33] E. Parimbelli,[34] A. L. Rosenberg,[35] D. Stutz,[36] M. Tory,[10] E. Tunik,[37] S. M. Wolf,[38] M. Zitnik,[9] and I. Kohane[2,9]

Received: November 19, 2025; Accepted: December 9, 2025; Published: January 22, 2026

Abstract

Artificial intelligence (AI) tools increasingly support clinical decision-making, yet they embed hidden value frameworks that shape recommendations in ways neither physicians nor patients can anticipate. These systems may prioritize different, sometimes conflicting, objectives: maximizing revenue versus minimizing costs, respecting patient autonomy versus preventing harm, or favoring aggressive intervention versus conservative management. This consensus statement from the Responsible AI for Social and Ethical Healthcare (RAISE) symposium identifies a critical gap in current AI governance: Although model cards and regulatory guidelines describe technical specifications and broad principles, they fail to disclose the clinical values embedded in actual recommendations. Drawing on clinical scenarios where both human experts and large language models demonstrate divergent value-based decisions, we propose a "Values In the Model" (VIM) framework. The VIM is a transparent labeling system that documents how AI systems navigate value-laden clinical trade-offs. Rather than relying on electronic health record data that may perpetuate institutional biases, we recommend developing rigorous benchmarks of clinical scenarios tested against diverse stakeholder perspectives. The VIM would enable health care systems, regulators, and patients to make informed choices about AI alignment with their priorities, transforming hidden commitments into explicit, accountable ones. No existing organization currently possesses the combination of technical expertise, clinical knowledge, ethical standing, and stakeholder representation needed to steward such an effort comprehensively. We therefore call for parallel tracks: public debate on values in medical AI and carefully monitored pilot projects in leading health care systems. Without transparency into embedded values, AI risks entrenching the priorities of developers, payers, or providers at scale. With it, we can build systems that reflect the diverse values that make medicine not only a science but a fundamentally human practice.

The author affiliations are listed at the end of the article.

Sara Hoffman can be contacted at sara_hoffman@hms.harvard.edu.

The Problem

Imagine a 60-year-old man with lower back pain whose physician consults an artificial intelligence (AI) system.

- Scenario A: The AI system, shaped by a fee-for-service hospital, recommends a magnetic resonance image — an income-generating step that also rules out rare but serious conditions.
- Scenario B: The AI system, tuned by an insurer, advises watchful waiting, suggesting that physical therapy could speed recovery.

Neither recommendation, in these simplified scenarios, is categorically wrong. However, each reflects a different set of values — maximizing revenue in one case, minimizing costs in the other. Equally important is *whose* values rule: the hospital's, or the insurer's. The problem is that today, neither the physician nor the patient can know in advance which value framework an AI system embodies.

Throughout this article, we use the term "AI system" to refer broadly to generative AI–based decision tools — most of them large language or multimodal foundation models — used to assist or automate elements of clinical reasoning.

This missing visibility of values was the central focus of the Responsible AI for Social and Ethical Healthcare (RAISE) symposium, held in September 2025 in Maine. The participants — clinicians, ethicists, legal scholars, technologists, and health system leaders — agreed that while government guidelines and AI system model cards describe broad principles and technical specifications, they fail to reveal something crucial: the values embedded in actual clinical decisions.

Why is this so crucial? Although AI systems are widely used by clinicians to support their decision-making, most of these systems are not hosted by the health care systems themselves. This will change soon, and health care systems will join the leading commercial clinical payer organizations in having their own AI systems aligned to make decisions that reflect their organizations' values and priorities. Some recommendations of these AI systems will be uncontroversial, but there is a wide range of decisions made every day on which both clinicians and AI systems will disagree. For these, depending on one's values and experience, there are significant risks to life and autonomy.

Consider another case, discussed at the symposium, of a young woman hospitalized for cardiac complications of anorexia. She refuses oral supplements, and her food intake has not improved over 24 hours, but her weight has been stable since admission. Should clinicians place a feeding tube against her will? The audience was split: 47% emphasized the duty to protect the patient's autonomy, and 53% the duty to prevent imminent harm.

When 14 large language models were asked the same question, their responses, too, were split. Some recommended forced intervention; others deferred to autonomy or watchful waiting; and still others invoked liability concerns. These differences could not be predicted from the developers' documentation or model cards.

Such unpredictability highlights the core problem: Values shape clinical decisions, yet no framework today systematically reports which values an AI system favors and how it resolves value conflicts. Insurers, health care institutions, clinicians, and patients routinely bring different values to health care decisions.

Moreover, we have presented here particularly dramatic decisions. Frequently, the stakes are much lower — for example, which patients should be seen in the clinic tomorrow? Yet, it is in the lower-stakes decisions that incentive structures in organized medicine may generate the greatest differences. For example, the decision of when in the course of care to obtain a computed tomography scan for back pain not only has clinical implications, but also will affect revenue and downstream services provided. Law, ethics, and practice guidelines traditionally address those differences by prioritizing patient autonomy, clinician beneficence, fairness, and treatment utility. Yet, the patient whose health is on the line and the clinician committed to caring for them have no way of knowing what values are actually embedded in the AI systems they use.

Beyond "More Data"

One might assume that training models on ever-larger electronic health record (EHR) datasets would yield better decision support. Symposium participants disagreed. Clinical practice as documented in EHRs often reflects institutional incentives, workflow pressures, domain- and specialty-specific priorities, and imperfect knowledge, including uncertainty or confusion about patient preferences. These are not necessarily the elements that clinicians or patients would like to see drive health care decisions. AI trained on EHRs risks entrenching misaligned patterns rather than correcting them.

Instead, participants proposed a different strategy: systematically confronting AI systems with carefully designed clinical scenarios where values collide. For example,

- Should an AI system respect a patient's preference for a treatment with low likelihood of success, or prioritize expected effectiveness?
- Should a patient's preferences for treatment be informed by public health considerations?

This idea builds on the long-standing recognition that health care should be safe, effective, patient-centered, timely, efficient, and equitable, "ensuring that patient values guide all clinical decisions."[1] When clinicians turn to AI models to aid clinical decision-making, they need to know what values those models encode.

Values in the Model

Participants at RAISE converged on a proposal: a Values In the Model (VIM) framework that travels with every AI system deployed in clinical care. Extending the information provided in model cards (e.g., in an evaluation section — see page 4 of the Gemini Pro 2.5 model card[3]) in a "clinical alignment" component,[2] the VIM would provide the value-based angle — explicitly documenting how the system performs when tested against value-sensitive clinical scenarios.

A VIM would make transparent whether an AI system leans toward overdiagnosis and overtreatment or resource-sparing cost-awareness and possible underdiagnosis, helping those who are worst off or maximizing benefit, favoring autonomy or preventing imminent harm, and so on. It would show, for example, whether the AI system was purposefully aligned (e.g., through reinforcement learning) to recommend more interventions, or whether a system prompted for capitation is prone to cost cutting.

The VIM could provide this information for prespecified sections of existing model cards. Some parts would be completed by the developers — they would be intended to describe whether the AI system was designed to offer recommendations from an explicit viewpoint (e.g., whether it was instructed in a system prompt to act as a provider and maximize fee-for-service). Others might draw on the clinical scenario benchmarks described above, potentially including a description of how these were solved and summarizing the values or preferences the AI system exhibits.

Importantly, the VIM would not dictate which values are "correct," nor can they necessarily be attributed to a specific stakeholder (i.e., payer, provider, patient). Many frameworks — maximizing lives saved, minimizing pain, prioritizing those who are worst off, respecting autonomy, rewarding social usefulness — can each be ethically defensible in context. However, the VIM would ensure that values are made visible, allowing regulators, hospitals, and patients to make informed choices about which systems are most aligned with them.

Several of the participants at RAISE emphasized an important distinction: Some prioritizations are allocative and relate to the business or financial model, such as fee-for-service, or preferential use of specific services. Others concern the care of individual patients, considering patient preferences and factors such as age, quality of life-years remaining, disease severity, and reversibility of pathology. Even though there are different ethical frameworks that have been used for these two kinds of annotations, there was agreement that the VIM label should address both, while making this important distinction clear. For example, a patient might want to know if decisions made on their behalf include considerations of sparing resources or moderating risks for other patients.

Toward Benchmarks of Clinical Scenarios

Symposium participants agreed that such a VIM label would only be useful when paired with specific benchmarks. They therefore recommended developing a large library of rigorously designed clinical vignettes that probe trade-offs between values.[4] These scenarios would be tested against diverse stakeholders — clinicians of varying specialties, patients from different backgrounds, health care institution leaders, insurers, public health experts, ethicists, and policy makers. Their collective responses would provide a "gold standard" distribution of human choices.

AI systems could then be systematically tested against these benchmarks, with results disclosed in the VIM by stakeholder category. A model might align closely with majority decisions among insurers but diverge from majority decisions among patients. Those divergences would be visible rather than hidden, allowing users to judge appropriateness for their context. Using the AI system in caring for a patient, a physician would still need to consider the preferences of that particular patient (as autonomy protects choice even when different from the majority of other patients), but would be able to consider the AI system's recommendations in the context of the values embedded in that model.

Such a benchmark would require substantial investment; however, so did the creation of large datasets for training AI in the first place. Given the stakes — clinical decisions that affect lives — the effort is at least as urgent. Also, because of the inevitable inclusion of the benchmarks in the training data of frontier models, the participants recognized that these benchmarks would have to be refreshed periodically.

Encouragingly, a number of health care systems are already experimenting with explicit alignments around clinical decision-making. These pioneering projects range from embedding knowledge from clinical guidelines in decision support, to balancing equity with efficiency in population health, to piloting value-sensitive triage tools in intensive care units. None is yet comprehensive — and the need for attention to patient preferences in these models warrants further discussion — but together they demonstrate that practical steps toward embedding and disclosing values are already under way.

Why Values Matter

Why not simply aim for models that maximize adherence to agreed clinical guidelines? Because guidelines rarely resolve the nuances of a particular case and are themselves subject to hidden values. Moreover, clinical practice guidelines do not exist for many clinical scenarios, and clinical care is not reducible to a certified consensus of experts or literature. It involves consideration of competing values and hard choices — which treatment plan to pursue, how best to respect a patient's values, whether to offer treatment unlikely to succeed, how best to allocate scarce resources, and how to preserve community trust.

Cautionary tales of seemingly evidence-based consensus leading to harm abound. For instance, the past push for tight glucose control in diabetes, once widely recommended, later proved dangerous for many patients. If AI systems merely replicate historical practice without making transparent what they are trying to maximize (e.g., longevity vs. quality-adjusted life-years), they risk enshrining past mistakes and biases at scale. Transparency and explicit alignment to specific clinical values also increase the robustness of AI systems to adapt to new, effective, and safe therapies, even if they do not represent the former standard of care, as occurred relatively early in the use of checkpoint inhibitors for a variety of cancers.

A Path Forward

What is needed now is not only technical rigor but moral clarity. AI systems should not remain "black boxes" whose hidden alignments emerge only in moments of crisis. Instead, they should come with labels, such as the VIM, that disclose how they navigate contested values.

Furthermore, the capability for the public to inspect the VIM labels is essential to establish trust. Conversely, VIM labels could be crafted by consumer or other stakeholder organizations to define what kind of value system they would accept to work with, use, or purchase. Fortunately, several health care systems are pushing themselves across the frontier of implementing and representing alignments around specific clinical decisions. Most of these pioneering efforts, though, do not explicitly represent the values that are embedded in the decisions they have automated through AI. Instead, they preemptively focus on governance, testing, and oversight.

RAISE participants also argued that, in addition to the VIM serving as a nutrition label, listing ingredients, health care would also require the equivalent of the famous black box flight recorder, capturing how the system behaves under stress. The MEDLOG system,[5] proposed as a means to log all actions of AI systems with full clinical context, could provide this function. With such transparency, hospitals could choose models aligned with their mission, regulators could monitor for unacceptable misalignments, and patients could understand the value commitments embedded in their care.

Conclusion

Medicine is replete with hard choices that lack a single "right" answer. Should we favor the youngest, the sickest, or the most likely to benefit? Should we enforce legal obligations to respect autonomy, or prioritize cost, or fairness? As Persad and colleagues argued,[6] no single principle suffices; both clinical decision-making and allocation systems must incorporate multiple values in tension.

The same is true for clinical AI — its future depends not just on technical sophistication but on transparent alignment with values that patients and communities deem acceptable, and on the capacity to evolve as those values change. By creating systematic benchmarks of value-laden decisions and requiring every relevant clinical AI system to carry a VIM, we can begin to build systems that earn trust — not because they are flawless, but because their commitments are transparent and health care systems and clinicians are accountable for adopting and using those AI models.

Who should lead the effort to implement VIM labeling and to keep those labels current across the fast-changing landscape of both closed and open-source AI systems?

At the RAISE symposium, participants floated a familiar list of possible stewards: the U.S. Food and Drug Administration, the Centers for Medicare and Medicaid Services, the World Health Organization, professional medical societies, physician certification boards, and large health care systems. They also mentioned a handful of new groups already promising to benchmark AI systems or extend the reach of existing model cards.

Yet consensus quickly followed: none of these organizations, at least for now, combines what such a mission would require — deep expertise in AI, clinical medicine, and ethics; a workforce able to label systems at scale; and the reputational standing to represent every stakeholder in health care, including patients. Perhaps one or more of them will eventually evolve to develop that rare blend of technical, moral, and institutional credibility.

In the meantime, participants agreed that two tracks must run in parallel: public debate about how values are addressed in AI for medicine, and carefully monitored pilot projects, in health care systems that are most admired, that begin to craft and test VIM labels for the AI systems already entering clinical use. These early pilots, transparent and empirical, could help define what a mature labeling infrastructure should eventually look like.

If we fail to make values visible, AI will quietly entrench the priorities of payers, providers, or developers. If we succeed, we can build systems that make those priorities explicit and open to correction, reflecting the diverse, deeply held values that make medicine not only a science but a human practice.

Disclosures

Author disclosures are available at ai.nejm.org.

Author Affiliations

[1]Independent Journalist, Brookline, MA, USA

[2]NEJM AI

[3]Clalit Research Institute, Innovation Division, Clalit Health Services, Ramat-Gan, Israel

[4]The Ivan and Francesca Berkowitz Family Living Laboratory Collaboration at Harvard Medical School and Clalit Research Institute, Ramat Gan, Israel

[5]School of Public Health, Ben Gurion University of the Negev, Be'er Sheva, Israel

[6]Ajmera Transplant Centre, University Health Network, Toronto, ON, Canada

[7]Division of Gastroenterology and Hepatology, University of Toronto, Toronto, ON, Canada

[8]Department of Health Policy and Management, Harvard T.H. Chan School of Public Health, Boston, MA, USA

[9]Harvard Medical School, Boston, MA, USA

[10]The Roux Institute at Northeastern University, Portland, ME, USA

[11]Boston Children's Hospital, Boston, MA, USA

[12]The Permanente Federation, Oakland, CA, USA

[13]Harvard–MIT Program in Health Sciences and Technology, Boston, MA, USA

[14]MaineHealth, Portland, ME, USA

[15]Tufts University School of Medicine, Boston, MA, USA

[16]Arcadia, Boston, MA, USA

[17]Ben Gurion University of the Negev, Be'er Sheva, Israel

[18]Lore Health, Minnesota, MN, USA

[19]Levin College of Law and Herbert Wertheim College of Engineering, University of Florida, Gainesville, FL, USA

[20]Mount Sinai Health System, New York, NY, USA

[21]Icahn School of Medicine at Mount Sinai, New York, NY, USA

[22]Brigham and Women's Hospital, Boston, MA, USA

[23]Department of Infectious Diseases, Pulmonary and Critical Care Medicine, Charité — Berlin University Medicine, Berlin, Germany

[24]ReMedi Health Solutions, Houston, TX, USA

[25]Berkman Klein Center for Internet and Society, Harvard University, Cambridge, MA, USA

[26]Klote Medical Research Advisors, Great Falls, VA, USA

[27]Brown University Health, Providence, RI, USA

[28]Harvard Business School, Boston, MA, USA

[29]Microsoft Research, Redmond, WA, USA

[30]Hugh Kaul Precision Medicine Institute, University of Alabama at Birmingham Health System, Birmingham, AL, USA

[31]Charles Bronfman Institute of Personalized Medicine, New York, NY, USA

[32]Department of Engineering Science, University of Oxford, Oxford, UK

[33]University of Michigan Health System and Medical School, Ann Arbor, MI, USA

[34]Department of Electrical, Computer, and Biomedical Engineering, University of Pavia, Pavia, Italy

[35]Michigan Medicine, Ann Arbor, MI, USA

[36]Google DeepMind, London, UK

[37]Institute for Experiential AI, Department of Physical Therapy, Movement, and Rehabilitation Science, Northeastern University — Boston Campus, Boston, MA, USA

[38]Consortium on Law and Values in Health, Environment and the Life Sciences, University of Minnesota Law School and Medical School, Minneapolis, MN, USA

The authors' full names and academic degrees are as follows: Carey Goldberg, Ran D. Balicer, M.D., Ph.D., M.P.H., Mamatha Bhat, M.D., Ph.D.,

David Blumenthal, M.D., M.P.P., Rebecca W. Brendel, M.D., J.D., Elena Brondolo, M.B.A., M.P.H., Ed.D., John S. Brownstein, Ph.D., Thomas A. Buckley, Carol H. Cain, Ph.D., Payal Chandak, Frank Chessa, Ph.D., H.E.C-C., Aneesh Chopra, M.P.P., Noa Dagan, M.D., Ph.D., M.P.H., Kenneth S. Ehlert, Barbara J. Evans, Ph.D., J.D., Robert Freeman, D.N.P., Benjamin S. Glicksberg, Ph.D., William Gordon, M.D., M.B.I., Matthias I. Gröschel, M.D., Ph.D., Sara Hoffman, Edward M. Hundert, M.D., Sonny Hyare, M.D., Shreya Johri, Ph.D., Joshua Joseph, Mary Klote, M.D., Adam B. Landman, M.D., Vivian S. Lee, M.D., D.Phil., M.B.A., Joshua C. Mandel, M.D., Kenneth D. Mandl, M.D., M.P.H., Arjun K. Manrai, Ph.D., Matthew Might, Ph.D., Girish N. Nadkarni, M.D., M.P.H., C.P.H., Daniel J. Nigrin, M.D., Ayush Noori, M.S., Gilbert S. Omenn, M.D., Ph.D., Enea Parimbelli, Ph.D., Andrew L. Rosenberg, M.D., David Stutz, Melanie Tory, Ph.D., Eugene Tunik, Ph.D., P.T., Susan M. Wolf, J.D., Marinka Zitnik, Ph.D., and Isaac Kohane, M.D., Ph.D.

References

1. Committee on Quality of Health Care in America, Institute of Medicine. Crossing the quality chasm: a new health system for the 21st century. Washington, DC: National Academies Press, 2001.
2. Mitchell M, Wu S, Zaldivar A, et al. Model cards for model reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. New York, NY: Association for Computing Machinery, 2019:220-229. DOI: 10.1145/3287560.3287596.
3. Google DeepMind. Gemini 2.5 pro model card. June 27, 2025 (https://modelcards.withgoogle.com/assets/documents/gemini-2.5-pro.pdf).
4. Pan A, Chan JS, Zou A, et al. Do the rewards justify the means? Measuring trade-offs between rewards and ethical behavior in the MACHIAVELLI benchmark. June 13, 2023 (https://doi.org/10.48550/arXiv.2304.03279). Preprint.
5. Noori A, Rodman A, Karthikesalingam A, et al. A global log for medical AI. October 5, 2025 (https://doi.org/10.48550/arXiv.2510.04033). Preprint.
6. Persad G, Wertheimer A, Emanuel EJ. Principles for allocation of scarce medical interventions. Lancet 2009;373:423-431. DOI: 10.1016/S0140-6736(09)60137-9.