

Medical Superintelligence: Foundations and Aspirations

Xiaoxuan Liu, Mayank Daswani, Christopher Kelly, Harsha Nori, Nicola Ding,
Matthew M Nour, Emma Rocheteau, Craig Sherstan, Viknesh Sounderajah,
Adrian D Haimovich, Eric Topol, Mustafa Suleyman, Dominic King, Eric Horvitz

Abstract

Advances in artificial intelligence (AI) have stimulated discussions about *medical superintelligence*, yet how to define and evaluate this aspiration remains unclear. We discuss how AI technologies can contribute to more effective care, introducing six interdependent dimensions of medical intelligence to frame contributions of AI to health care delivery. As AI capabilities and their integration into clinical workflows advance, crossing the threshold to medical superintelligence will require more than superhuman performance on isolated tasks. It will require systems that harness AI, independently and in collaboration with humans, to surpass today's strongest human-led health systems in integrated medical reasoning across the full arc of care. Realising their value will require understanding how these capabilities complement existing medical intelligence and combining them effectively to improve the quality and accessibility of care.

Introduction

Advances in AI invite us to consider how new capabilities could improve the quality and availability of health care, and what it would mean to achieve *medical superintelligence*. To frame these aspirations, we begin with the intrinsic intelligence underlying today's health care. We view *medical intelligence* as the capacity to integrate medical knowledge, reasoning, longitudinal context, care coordination, and human values into an evolving understanding of health and illness and to translate those insights into actions that improve the health and well-being of individuals and populations. This capacity can be exercised through individual clinicians and teams, health care organizations, AI systems, and human-AI collaborations.

We define *medical superintelligence* as the capacity to surpass the strongest contemporary human-led health systems through integrated medical reasoning across the full arc of care. Understanding the core constituents of medical intelligence, and how they work together, is essential to identifying how AI can enable such progress. We distinguish medical intelligence conceptually from clinical excellence, which represents how effectively available intelligence, resources, and workflows are brought together to deliver the best possible care.

To gauge progress, we must look beyond benchmarks of individual tasks, such as answering medical examination questions or identifying diagnoses from curated patient vignettes, and towards more realistic, clinical tasks involving continuous patient trajectories, meaningful outcome targets, and co-ordination among existing components of the health system.[1-3] Whilst benchmarking efforts are important, current benchmarking tasks alone cannot establish influences of AI on broader medical intelligence of health care delivery. AI's impact will ultimately depend on how effectively its capabilities translate into better, more accessible, and more efficient care through integration with clinicians and health care systems.

Two questions are central: what underlying intelligence capabilities are required, and how must they work together to support increasingly comprehensive and effective care? Progress beyond the intelligence underpinning the strongest health systems of today may require coordinated networks of AI subsystems that combine complementary expertise and act across settings and time.[4] Thus, we should assess an AI system's contribution to medical intelligence not only by its standalone performance, but also by how its integration into the broader system of care strengthens health care delivery. As these capabilities advance, coordination itself may increasingly be performed by AI, with humans providing complementary expertise, guidance, and oversight. This coordination should be reflected in our evaluation approaches, including outcome-based evaluations involving direct comparison with leading expert teams and health systems.[5] Evaluation should include stress testing under incomplete, conflicting, or misleading information; assessment across patient populations and care settings[6]; and prospective studies in

real clinical workflows. We note that evaluation becomes increasingly difficult as we attempt to measure intelligence superior to the current gold standard. For example, when judged against expert clinician consensus, superior reasoning may initially appear to be error than improvement. Ultimately, demonstrating significant gains will require evidence of improved downstream outcomes from interventional studies.

We propose six dimensions of medical intelligence as guideposts for developing, evaluating, and integrating AI systems into care. These dimensions can be realised within individual systems or through the coordinated contributions of clinicians, teams, and AI technologies.

1. Medical Knowledge and Its Application

A central competency of medical intelligence is the ability to access, evaluate, and apply medical knowledge of sufficient breadth and depth. Rather than functioning as a static library of facts, the system must maintain up-to-date knowledge grounded in the latest evidence. Applying this knowledge requires causal reasoning in individual cases: linking underlying biological mechanisms to observed manifestations, drawing on biological principles, predicting pathophysiological trajectories, and formulating testable hypotheses. A medically intelligent system might surpass existing knowledge, learning structured latent representations of biology and disease from associations across vast amounts of data, capturing unknown relationships among molecular mechanisms, physiological processes, clinical phenotypes, treatments, and outcomes. These representations provide a foundation for clinical reasoning in individual cases. A patient's state, trajectory, and response to treatment can be represented and inferred within this learned structure, allowing the system to anticipate how illness may evolve, and adapt its testing and therapeutic recommendations as circumstances change. Such learned representations may encode structure not yet articulated in the literature; they can also surface previously unrecognized relationships between entities, supporting the identification of rare diseases, hypotheses about undescribed mechanisms, and the recognition of new syndromes or disease entities.

To produce beneficial outcomes, knowledge must be integrated across complex, concurrent, interacting pathophysiological processes. Patients commonly present with coexisting conditions; medical intelligence therefore requires constructing a coherent model of the patient's overall clinical state. This integration of information needs to traverse multiple modalities, including speech, observable physical manifestations, biological samples, medical images and physiological signals. A medically intelligent system should be capable of interpreting and integrating different modes of medical information, understanding how they relate to each other. It should be robust to surface-level variations, whilst also sensitive to deviations from an individual's baseline or known

patterns of human health, and use mechanistic reasoning to develop and test hypotheses about the underlying pathology even when formal diagnostic labels are absent from existing knowledge. Such multimodal intelligence depends on reliable physical sensors, while continuous biosensors increasingly provide granular day-to-day data points to inform reasoning about deviations from an individual's baseline.[7]

When devising a treatment plan, the system must anticipate and minimise the risk that interventions targeting one condition will worsen another. Multi-medication management requires modeling individualized drug effects, interactions, dosing requirements, and therapeutic windows, seeking to maximize efficacy while minimising adverse effects. Doing so requires integrating knowledge of pharmacokinetics, pharmacodynamics, physiological variability, and drug-drug interactions.

Because medicine evolves, a medically intelligent system must be capable of self-directed learning. It should continuously acquire, evaluate and consolidate new information into its knowledge base whilst operating with epistemic humility: explicitly representing the limits of current medical knowledge and its own knowledge, and knowing when to seek additional information.

A medically intelligent system should be able to evaluate and integrate evidence-informed perspectives across traditions and cultural contexts while accounting for variation among populations and resource availability. It should recognize that relevant knowledge may be underrepresented in the dominant medical literature, including insights arising from specific populations, care traditions, and clinical contexts. Such knowledge should be evaluated according to empirical standards appropriate to the claim, with causal claims requiring corresponding evidentiary support.

2. Diagnostic Reasoning and Triage

Diagnostic reasoning requires a system to move through a diagnostic pathway with probabilistic rigor while weighing the value, cost, and burden of acquiring additional information. This entails Bayesian reasoning: maintaining a prior distribution over possible diagnoses informed by encoded medical knowledge, and continuously updating diagnostic probabilities as new information becomes available.[8-10] Triage is a related application of this capability and an important use of medical AI. The core task of triage is to identify an appropriate next action and its urgency under incomplete information, balancing likely causes of symptoms against low-probability, high-consequence conditions. Triage should recommend an appropriate level and setting of care, urgency of assessment, and conditions that should prompt reassessment or escalation, while accounting for the range of trajectories that may follow from a single snapshot in time.

When the goal is diagnosis, the system must assess the availability and expected value of information provided by diagnostic tests, weighing that value against financial costs, resource use, invasiveness, and physical or psychological burden on the patient. The system should preserve, and appropriately represent, diagnostic uncertainty rather than prematurely closing the diagnostic process. As evidence accumulates over the sequential diagnostic process, prior hypotheses and decisions must often be revisited and revised. Given that diagnostic and therapeutic errors may occur, error recognition and recovery are themselves core capabilities. A medically intelligent system should identify prior errors, incorporate newly available information, and revise hypotheses, diagnoses, treatment, and prevention plans accordingly.

Particularly challenging cases include rare, atypical, or previously uncharacterised conditions. Challenges can also arise when patients express symptoms in atypical or incomplete ways, for example, due to low health literacy or emotional distress.

3. Therapeutic Decision-Making and Action

Treatment planning requires the system to account for individual characteristics and evolving status, adapting decisions in a closed loop as treatment responses become apparent. For example, when initiating antihypertensive therapy, the system should use serial blood pressure measurements, adverse effects, laboratory findings, the individual's lifestyle and routines, and patient-reported outcomes to guide the therapeutic approach, including drug selection, dose adjustment and lifestyle modification at clinically appropriate intervals. A medically intelligent model can combine the full patient context, all available test results and its existing knowledge base of population health trajectories, to recommend effective treatments and anticipate individual responses while accounting for uncertainty.

In the real world, treatment decisions are often required despite incomplete or conflicting information, and sometimes in the absence of a definitive diagnosis, to alleviate symptoms or reduce the risk of irreversible harm. For treatment efficacy, well-conducted randomised trials and their systematic reviews generally provide the strongest evidence, but applying their findings to an individual patient requires assessing differences between that patient and the populations studied. When deliberating about treatments, the comparative benefits, harms, costs, and local availability of alternative therapies must be weighed carefully. Therapeutic plans must also account for resources required for monitoring and follow-up.

Beyond treatment planning, interaction between intelligent systems (humans, instruments and robotic systems) and the physical world will enable more precise and sophisticated interventions on the human body. Multisensory perception as a component of medical intelligence will be essential for executing physical interventions such as

surgery and other procedures, where anatomical knowledge and operative planning must be integrated with visual and tactile feedback, control of movement, and adaptation to unexpected findings.

All treatment plans should be aligned with the individual's goals, values and preferences. Personalization requires understanding how individual preferences align with or diverge from standard clinical recommendations and assessing the consequences of such divergence. This involves integrating information about the person's lifestyle, functional abilities, treatment burden, access to health care, prior experiences, and expressed preferences. A medically intelligent system should use this understanding to develop clinically appropriate plans that prioritize outcomes consistent with individual values.

4. Longitudinal Context

The preceding dimensions depend on information accumulated over time. Longitudinal context addresses the capacity to preserve, interpret, and use that information across the course of care. Medicine today is often fragmented across discrete encounters: clinic visits, acute care episodes, and inpatient admissions. Each encounter may address a specific problem within one part of the health care system, with incomplete access to the patient's broader history and ongoing care. A medically intelligent system should possess longitudinal memory, enabling it to integrate prior health events into ongoing inference and decision-making. Rather than merely recalling events, it should distinguish durable facts from transient observations, preserve their temporal relationships, and update its model of the individual as circumstances change. As updates occur, the evolving model should retain attributional connection with the original information sources and continue to update dynamically as new information arrives.

Long-term context can support proactive clinical stewardship and prevention, which current models of care often struggle to deliver consistently. The system should observe and model a patient's health trajectory over extended periods, identifying baseline patterns across multiple dimensions, factors associated with better health including psychosocial and socioeconomic factors, and early signals of deterioration.[11-13] Subtle trends, such as gradual changes in biomarkers, changes in functional status, patterns in biosensor measurements, or evolving day-to-day symptoms, are precisely the signals that episodic, reactive care often fails to detect.

Maintaining high-fidelity longitudinal context raises important issues involving information provenance, record management, and privacy. Beyond retention, longitudinal memory requires effective context-sensitive search and retrieval, as well as temporal abstraction that preserves and summarizes salient information. It must also retain uncertainty and source attribution, reconcile conflicting records, and prevent outdated or erroneous information from propagating into subsequent decisions.

5. Clinical Coordination

Effective care requires the orchestration of capabilities that no single clinician or AI system can exercise alone, including coordinating distributed intelligence, action, and resources across the broader care ecosystem. Whereas longitudinal context integrates information across time, clinical coordination integrates expertise and action across people, systems, and organizational structures. The design and deployment of medical AI therefore require close attention to how AI systems coordinate with clinicians and operate within the workflows and constraints of health care organizations.

A key capability of a medically intelligent system is recognizing its own limitations and acting appropriately by referring to, collaborating with, and incorporating external sources of expertise, tools, and resources. Even a system with superhuman medical reasoning would operate in a world of distributed information, physical resources, specialized capabilities, institutional authority, and human relationships. Therefore, this capability requires an understanding of the system's own limitations across medical knowledge, diagnostic reasoning, and treatment planning, as well as awareness of available resources and the complementary strengths of human and AI collaborators. It also requires determining who is best placed to act, when responsibility should transfer, and how to verify that agreed actions and follow-up have occurred. Thus, this dimension of medical intelligence is inherently *relational*: it governs how individuals and teams situate their abilities within a broader ecosystem of medical expertise rather than attempting to operate as closed systems.

6. Alignment with Human Values and Agency

Medical intelligence must recognize the central role of patients' experiences, values, and agency in health care. Beyond clinical reasoning and coordination across the medical ecosystem, advanced medical intelligence must engage with the fundamentally human dimensions of care.[14] Medicine is not solely the optimization of biological systems; it is the care of people whose experiences, values, relationships, and preferences shape both the meaning of illness and the goals of treatment.

At the core of this capability is the capacity to understand another person's perspective and its relevance to clinical care: how patients perceive, interpret, and experience their condition. This includes recognizing how symptoms are experienced and communicated, how individuals form beliefs about their health, and how fears, hopes, and prior experiences influence their decisions over the course of care.

Such understanding must extend beyond the individual to encompass the broader social context, including family members, caregivers, support networks and other relationships that often play central roles in care decisions. A medically intelligent system must be able

to reason about values and tradeoffs while supporting individuals as the principal agents in decisions about their health. Accordingly, the system must be able to elicit, represent, and incorporate individual values into decision-making, including how people grapple with uncertainty, and weigh quality of life, health risks, functional outcomes, and short-versus long-term goals.

To deliver human-centered medical intelligence, the system must communicate with empathy and clarity while maintaining appropriate professional and social boundaries. Effective care depends not only on accurate reasoning, but also on earning appropriately calibrated trust, conveying complex information clearly and transparently, and supporting patients through uncertainty. The “human touch” in medicine, including presence, availability, and responsiveness, remains a critical component of high-quality care.

In individual care, the system should not impose preferences but rather elicit patients’ values and support reasoning within their values framework. This includes navigating tensions between autonomy and beneficence, reasoning through ethical conflicts, and making explicit the assumptions and value judgments behind recommendations on different courses of therapy. Respect for agency includes supporting informed refusal while clearly explaining consequences and offering alternative options.

At the population level, improving aggregate health may require choices about the distribution of benefits, burdens, and risks. Calls for medically intelligent systems to make values trade-offs explicit have been published.[15] Transparency about how the system responds to value-sensitive decisions will be essential for assessing its effects on different groups, and supporting accountable decision making under clearly stated ethical and policy priorities.

Integrative Intelligence

The six dimensions described above collectively define the constituent capabilities of medical intelligence. Their integration gives rise to *integrative medical intelligence*: the capacity to bring these capabilities together in coherent clinical reasoning and action. While each dimension can be described independently, real-world medical care requires continuous integration rather than their isolated exercise.

Integration is not merely additive. Decisions in one domain alter the reasoning required in others: new diagnostic information changes therapeutic options; treatment response changes diagnostic beliefs; longitudinal context changes the interpretation of current evidence; the availability of specialist care changes which plans can be carried out; and patient values change which outcomes should be pursued. Integrative intelligence is therefore the capacity to manage these dependencies as a coherent whole.

Integration also creates opportunities for collective learning. An integrated system could learn collectively from outcomes, errors, and feedback across its deployments. Experience accumulated across many patients and care settings could inform shared updates, allowing lessons from one setting to improve performance elsewhere. This could increase both the scale of learning and the speed at which improvements are disseminated.

Conclusion

We proposed six interdependent dimensions as a framework for understanding the medical intelligence underlying today's health care and identifying how AI can strengthen it. These dimensions provide a basis for characterizing strengths, limitations, and progress across health professionals, health systems, AI systems, and human-AI teams. We offer this framework as a starting point for discussion and welcome extensions and perspectives not captured here. Although separated here for analytical clarity, the dimensions are interdependent: the hallmark of advanced medical intelligence is not excellence along any single axis, but coherent performance across them. Progress towards medical superintelligence will require advancing both these capabilities and their integration. The threshold is the capacity to surpass today's strongest human-led health systems in integrated medical reasoning across the full arc of care, whether AI contributes independently or in collaboration with people.

Realizing the benefits of medical intelligence will require integrating these capabilities with future medical practice. Highly capable systems must operate within clinical, ethical, regulatory, economic, and institutional constraints. To understand frontier capabilities, however, it is important to distinguish underlying intelligence from readiness for safe and effective deployment. Evidence of these capabilities should also be distinguished from evidence that their deployment improves care. Transformative technologies often outpace the structures designed to govern them; understanding the extent and limitations of their capabilities is therefore essential to determining how they should be effectively and responsibly deployed. A system may possess capabilities that cannot yet be safely, legally, economically, or institutionally deployed; conversely, current barriers to deployment do not define the limits of the underlying intelligence. This distinction is especially important because frontier capabilities may enable workflows, roles, and models of care that do not yet exist and whose implications may become clear only through experimentation, carefully evaluated deployment, and the longer-term co-evolution of health care practices and systems. It is possible that realising the full benefits of medical superintelligence will require broader restructuring of health care. As we pursue these changes, the goal must be to improve the quality and accessibility of care, and thereby health and wellbeing, while preserving humanity, connection, and compassion as essential qualities of care.

Funding: Microsoft provided salary support to the authors it employs. No additional funding was received for this article.

Competing interests: XL, MD, CK, ND, VS, HN, DK, EH, MS, ER, MN and CS are employees of Microsoft. MD, CK, ND and VS hold shares in Google. CS holds shares in Sony. AH is a consultant for MBO Partners. ET is an adviser to Microsoft and receives consulting fees from Abridge and Mercor. No other competing interests were declared.

Ethics approval: Not applicable. This article does not report original research involving human participants or animals and contains no identifiable patient data.

References

1. Goh E, Wu D, Walton C, et al. Toward a test of medical AI superintelligence. *Nat Med* 2026;32:3128-30. doi:10.1038/s41591-026-04539-8.
2. Gallifant J, Bitterman DS. Humanity's next medical exam: preparing to evaluate superhuman systems. *NEJM AI* 2025;2(11). doi:10.1056/AIe2501008.
3. Cresswell K, Williams R. How artificial intelligence is disrupting the field of medical knowledge and evaluation: a narrative review. *BMJ Digit Health* 2026;2:e000104. doi:10.1136/bmjdh-2026-000104.
4. Genewein T, Franklin M, Lerchner A, et al. From AGI to ASI. arXiv 2606.12683v1 [Preprint]. 10 June 2026 [cited 9 October 2026]. doi:10.48550/arXiv.2606.12683.
5. Bedi S, Cui H, Fuentes M, et al. Holistic evaluation of large language models for medical tasks with MedHELM. *Nat Med* 2026;32:943-51. doi:10.1038/s41591-025-04151-2.
6. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat Med* 2024;30:2613-22. doi:10.1038/s41591-024-03097-1.
7. Perez MV, Mahaffey KW, Hedlin H, et al. Large-scale assessment of a smartwatch to identify atrial fibrillation. *N Engl J Med* 2019;381:1909-17. doi:10.1056/NEJMoa1901183.
8. Nori H, Daswani M, Kelly C, et al. Sequential diagnosis with language models. arXiv 2506.22405v1 [Preprint]. 27 June 2025 [cited 9 October 2026]. doi:10.48550/arXiv.2506.22405.
9. McDuff D, Schaekermann M, Tu T, et al. Towards accurate differential diagnosis with large language models. *Nature* 2025;642:451-7. doi:10.1038/s41586-025-08869-4.
10. Tu T, Schaekermann M, Palepu A, et al. Towards conversational diagnostic artificial intelligence. *Nature* 2025;642:442-50. doi:10.1038/s41586-025-08866-7.
11. Shmatko A, Jung AW, Gaurav K, et al. Learning the natural history of human disease with generative transformers. *Nature* 2025;647:248-56. doi:10.1038/s41586-025-09529-3.
12. Kraljevic Z, Bean D, Shek A, et al. Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. *Lancet Digit Health* 2024;6:e281-90. doi:10.1016/S2589-7500(24)00025-6.

13. Waxler S, Blazek P, White D, et al. Generative medical event models improve with scale. arXiv 2508.12104v1 [Preprint]. 16 August 2025 [cited 9 October 2026]. doi:10.48550/arXiv.2508.12104.
14. National Academy of Medicine. Artificial intelligence code of conduct. <https://nam.edu/programs/artificial-intelligence-code-of-conduct/> (accessed 14 Oct 2026).
15. Goldberg C, Balicer RD, Bhat M, et al. The missing dimension in clinical AI: making hidden values visible. NEJM AI 2026;3(2). doi:10.1056/AIp2501266.