



FUNDAMENTALS OF MEDICAL ETHICS

The Ethics of Relational AI — Expanding and Implementing the Belmont Principles

Ida Sim, M.D., Ph.D., and Christine Cassel, M.D.

November 2022 marked an inflection point in the public's experience with artificial intelligence (AI), when ChatGPT, a “large language model” (LLM), chatted with millions of

people in everyday English. Previously, AI had primarily taken the form of predictive analytics that produced numerical predictions — for example, “this patient has a 68% chance of requiring intensive care within the next 24 hours.” In contrast, LLMs belong to a new class of AI called generative AI (GenAI) and generate natural language, the medium in which humans communicate with each other. GenAI is thus relational AI, which is qualitatively different from predictive AI. Inasmuch as “the medium is the message,” GenAI's relational nature raises additional ethical questions beyond the already-daunting ethics surrounding predictive AI.

The 1979 Belmont Report established basic ethical principles for research involving humans: beneficence, respect for persons, and justice. These principles are also often applied to clinical care. Traditionally, physicians held fiduciary responsibility for upholding these principles in the best interests of patients. The ethical calculus changes, however, when AI-generated text, speech, images, and video are interposed between clinicians and patients: in these situations, clinicians themselves are subject to AI and therefore also deserve beneficence, respect, and justice.

There are countless existing and potential applications of re-

lational AI throughout health care. Examples include GenAI drafting patient-portal messages or care-handoff summaries for trainees; conversational interfaces for patients to learn about their diagnosis, consider treatment choices, or prepare for surgery, all at their literacy level in their own language; and ChatGPT-like interfaces allowing patients to self-diagnose and access treatment advice. Soon, convincingly lifelike avatars — perhaps of a patient's own clinician — will replace today's text-prompt interfaces. GenAI's potential to ubiquitously supplement or replace human-mediated health care interactions, for good or ill, increases the necessity of an effective practical ethical response.

A plethora of ethical frameworks, guidelines, and principles exist for health AI (see the Supplementary Appendix, available at NEJM.org). These initiatives are

remarkably consistent in calling for AI that is fair, appropriate, valid, effective, and safe (FAVES),¹ as well as transparent, explainable, and inspectable. These principles track with the Belmont principles, but the complexity of AI, especially relational AI, makes it hard-

was in 2020). The Coalition for Health AI — a public-private partnership involving academia, technology companies, and the federal government — has proposed development of a national network of health AI assurance laboratories to evaluate the safety

AI has the potential for such broad societal implications,³ we believe the concept of health AI beneficence must be expanded to include beneficence to communities and to society generally.⁴

Second, aiming at respect for persons, most published health AI guidance documents feature “transparency.” The AI principles outlined by the U.S. Department of Health and Human Services (HHS) distinguish three types: transparency about how patient data are being used, clarity about the role AI systems are playing in decision making, and allowing regulators or overseers access to the AI algorithms themselves.¹ Such transparency, along with patients’ ability to decline the use of AI, could be equated with seeking informed consent and thus reflects respect for patients as persons.

What about clinicians? If they are respected as persons, clinicians also deserve transparency and the right of refusal regarding the ways that data on their practice patterns are used for AI model building, the basis on which AI systems determine treatment recommendations and, crucially, whether AI-generated avatars may impersonate clinicians in interactions with patients. Relational AI raises a deeper question of the meaning of “respect for persons.” A broader notion of respect as the “recognition of the unconditional value of patients as persons”⁵ may offer a starting point, but the concept should be expanded further to include recognition of clinicians as persons.

Third, “fairness,” the first HHS FAVES principle,¹ is an expression of justice, which dictates that AI benefits should be distributed equitably among populations and individuals. Technical mechanisms

Left on their own, some GenAI systems may pursue goal-oriented behavior that is misaligned with medicine’s moral tenets (for example, learning new ways to convince clinicians to provide treatment that benefits payers rather than patients).

er to translate those principles into effective ethical oversight.

First, the calls for AI to be appropriate, valid, effective, and safe fall under the Belmont principle of beneficence. Although technical methods exist for reducing the risk and impact of GenAI “hallucinations” (e.g., fabricated citations), LLMs pose an inherent risk of harm, intentional or otherwise, due to the probabilistic way in which they generate content. In addition, errors can arise from biased or erroneous training data and from deliberate misuse (e.g., misinformation, image manipulation). Moreover, unlike a drug whose chemical structure does not change over time, an AI’s performance can deteriorate (or “drift”) as soon as it is deployed, for reasons that may be planned (such as model updates) or unplanned (such as changes in the underlying statistical relationship of variables; e.g., anosmia is a rarer Covid symptom in 2024 than it

and effectiveness of AI in centralized settings using representative data sets.² Centralized evaluation is akin to factory testing and should be complemented with vigilance and ongoing point-of-care monitoring of health, workforce, and health system outcomes to ensure continued beneficence and fairness.


Some AI ethics frameworks also stress “appropriateness,” which falls within the concept of beneficence. The issues here are similar to those arising in clinical care rationing, in which conflicting priorities, structural biases, and philosophical questions of comparative utility collide. But GenAI presents an extra twist. Left on their own, some GenAI systems may pursue goal-oriented behavior that is misaligned with medicine’s moral tenets (for example, learning new ways to convince clinicians to provide treatment that benefits payers rather than patients). Because relational

for enhancing fairness include using representative data sets, including data on factors such as social determinants of health to identify and manage structural societal biases, and being transparent and explicit about the AI system's goals. There will always be statistical bias, no matter how large or "representative" a data set is, so the objective should be to understand and manage that bias, not to eliminate it entirely.

Algorithmic fairness is only one contributor to the overall fairness of an AI system. An algorithm that is theoretically fair (one that provides the same treatment recommendation for all patients with the same clinical features, for example) can be unfair in practice if it's deployed only by some clinicians for some patients. Relational AI introduces additional risks of unfairness if AI-generated language is used to manipulate humans into acting unfairly.

Fairness should also be considered in the distribution of any commercial or efficiency benefits. GenAI has the potential to improve the skills and efficiency of health care workers — and to eliminate the need for whole classes of workers. Will efficiency gains be distributed fairly? If patient data or practice-pattern data are used to train commercial models, should the profits be shared with the patients and clinicians who contributed those data?

These challenging ethical questions go beyond the traditional application of the Belmont principles.

 **An audio interview with Ida Sim is available at NEJM.org**



An emerging risk is the exacerbation of physician burnout owing to increased moral injury if AI is used to undermine physicians' professional role as fiduciaries for pa-

tients' best interests. The medical profession would be remiss if it narrowly restricted the consideration of health AI ethics to the protection of patients and their health outcomes.

The core Belmont principles are proving flexible enough to provide the conceptual framework even for such a radically new technology as GenAI. Within this framework, however, further explication of beneficence, respect for persons, and justice is needed, as is expansion of scope to include clinicians and the broader society. Implementation of these principles will require focused guidelines and codes of conduct from governmental, academic, professional, and industry groups to ensure that clinical care accords with the principles. But AI is being developed and deployed far too rapidly for the traditional approach of years-long ethical consensus development. Health AI guidelines and codes of conduct will have to continually evolve to keep pace with rapid advances in AI and technology.

We therefore call for the creation of a national network of proactive health AI ethics centers that would develop foundational methods; create and maintain living guidelines; coordinate the development, testing, and evaluation of implementation strategies far beyond institutional review boards; monitor ethics implementation as it evolves; and identify and share best practices. These centers should include experts in AI-driven health care, bioethics, human-factors engineering, implementation science, philosophy, and law. They should also foster full and active participation by patients and clinicians, using strong community partnerships and platforms to ensure input from all

communities including those that are marginalized. Funding of bioethics research should be bolstered by existing agencies such as the National Institutes of Health and the National Science Foundation. Additional funding might also be obtained from incentive programs of the Centers for Medicare and Medicaid Services or by other mechanisms to enable participation and buy-in from health systems.

Health AI carries no less potential for harming persons and society than do pharmaceutical products or medical devices. Yet only the European Union's AI Act, enacted in December 2023, has established federal-level regulatory power, backed by law, over AI. In the United States, the Food and Drug Administration, other state and federal agencies, the Coalition for Health AI, and others are developing a regulatory regime for health AI safety and performance. Drawing on different expertise and with a different type of accountability, a separate authority is needed to ensure that health AI adheres to the principles of beneficence, respect for persons, and justice. Such an entity might resemble a National Transportation Safety Board, but it is essential that it be proactive rather than reactive. We believe this authority should regularly issue short, actionable reports to guide the moral adoption of AI technologies in whatever form they may take in the coming years, and to report transparently on the state of health AI ethics to build public trust.

We are at a threshold moment for shaping the nature of medical care in the near and distant future. The most crucial choices confronting the medical community are

not technical but ethical: the paramount fiduciary responsibility is to all humans and to the persistence of medicine as a moral and human profession.

The series editors are Bernard Lo, M.D., Debra Malina, Ph.D., Geneva Pittman, M.P.H., and Stephen Morrissey, Ph.D.

Disclosure forms provided by the authors are available at NEJM.org.

From the Department of Medicine (I.S., C.C.) and the Program in Computational

Precision Health (I.S.), University of California, San Francisco, San Francisco.

This article was published on July 13, 2024, at NEJM.org.

1. Department of Health and Human Services. Health data, technology, and interoperability: certification programs updates, algorithm transparency, and information sharing, final rule. *Fed Regist* 2024;89(6):1192-438 (<https://www.govinfo.gov/content/pkg/FR-2024-01-09/pdf/2023-28857.pdf>).
2. Shah NH, Halamka JD, Saria S, et al. A nationwide network of health AI assurance laboratories. *JAMA* 2024;331:245-9.

3. Russell S. *Human compatible: artificial intelligence and the problem of control*. New York: Penguin Random House, 2020.

4. McDermott R, Hatemi PK. Ethics in field experimentation: a call to establish new standards to protect the public from unwanted manipulation and real harms. *Proc Natl Acad Sci U S A* 2020;117:30014-21.

5. Beach MC, Duggan PS, Cassel CK, Geller G. What does 'respect' mean? Exploring the moral obligation of health professionals to respect patients. *J Gen Intern Med* 2007;22:692-5.

DOI: 10.1056/NEJMp2314771

Copyright © 2024 Massachusetts Medical Society.

Anticipating the Next Pandemic

H. Cody Meissner, M.D., Bill G. Kapogiannis, M.D., and Daniel N. Wolfe, Ph.D.

Reflecting on a pandemic that killed nearly twice as many people in the United States as the 1918 influenza pandemic — which had been the worst infectious disease outbreak in the country's recorded history before Covid-19 — offers several lessons for pandemic preparedness and response. First, the Covid-19 pandemic led to extraordinary advances in vaccinology, resulting in the availability of safe and effective vaccines and demonstrating the ability of the medical community to rapidly address a major challenge in the face of an urgent public health need. Paradoxically, a second lesson is about the fragile state of the national and global vaccine enterprise, including issues associated with vaccine distribution and acceptance. A third lesson is that partnerships involving private, government, and academic resources were critical for facilitating the rapid development of the first generation of Covid-19 vaccines. Building on these lessons in the current in-

terpandemic period, the Biomedical Advanced Research and Development Authority (BARDA) is seeking to support the development of a new generation of improved vaccines.

Project NextGen is a \$5 billion initiative sponsored by the Department of Health and Human Services aimed at developing next-generation medical countermeasures against Covid-19.¹ This initiative will support double-blind, active-comparator, controlled phase 2b trials assessing the safety, efficacy, and immunogenicity of experimental vaccines relative to approved vaccines in ethnically and racially diverse populations. We anticipate that the vaccine platforms could be adapted to vaccines for other infectious diseases, thereby enabling a rapid response to future health security threats. These trials will address several considerations (see table).

The primary end point of the proposed phase 2b clinical trials will be a greater than 30% improvement in vaccine efficacy over

a 12-month period relative to currently approved vaccines. Efficacy will be based on protection against symptomatic Covid-19; in addition, participants will conduct weekly self-testing using nasal swabs to capture data on asymptomatic infections as a secondary end point. Whereas vaccines that are currently available in the United States are based on the spike antigen and are delivered intramuscularly, next-generation candidates will rely on more diverse platforms that contain spike genes as well as more conserved regions of the viral genome, such as the genes encoding nucleocapsid, membrane, or other nonstructural proteins. New platforms may include recombinant viral vector vaccines that use replication-incompetent or -competent vectors and contain genes encoding SARS-CoV-2 structural and nonstructural proteins. Second-generation, self-amplifying mRNA (samRNA) vaccines are a rapidly emerging form of technology that will be evaluated as another option. SamRNA