REVIEW ARTICLE

**AI IN MEDICINE**
Jeffrey M. Drazen, M.D., *Editor*

# Medical Artificial Intelligence and Human Values

Kun-Hsing Yu, M.D., Ph.D., Elizabeth Healey, S.B., Tze-Yun Leong, Ph.D., Isaac S. Kohane, M.D., Ph.D., and Arjun K. Manrai, Ph.D.

Recent progress in generative artificial intelligence (AI) has given rise to large language models (LLMs) that can be prompted to craft persuasive essays,[1] pass professional competency examinations,[2-4] and write patient-friendly empathetic messages.[5] Amid growing recognition of the capabilities of LLMs, many people have expressed concerns about their use in medicine and health care, citing known risks of confabulation, fragility, and factual inaccuracy.[6] As these risks are measured and mitigated, some of the unresolved questions that are coming into focus concern the "human values" that will remain embedded in AI models, both in their creation and in their use, and how the "values of an LLM" may not align with human values even if LLMs no longer confabulate and have been scrubbed of obvious toxic output. Such human values pertain broadly to the principles, standards, and preferences that reflect human goals and guide human behaviors (see the Glossary). As we review here, LLMs and new foundation models, as technically impressive as they are, are only the latest incarnation in a long line of probabilistic models that have been integrated into medical decision making, which have all required that their creators and implementers make value judgments.

Many of the challenges we address here were evident to the pioneers of medical decision analysis of the 1950s[7] and to scholars in subsequent decades[8-11] who conducted careful and creative studies of both human and algorithmic decision making to disentangle probability (i.e., the chance of an event occurring) from utilities (i.e., the quantified value judgments that are often only indirectly articulated in much of medical decision making). The nuanced understanding of individual values and risks is what makes the thoughtful clinician so indispensable. These considerations have renewed relevance with unprecedented and ubiquitous AI models such as LLMs. In this article, we first describe how value judgments enter predictive models in the context of familiar clinical equations and new AI language models. We then connect early work in reasoning about probabilities and utilities to the emerging issues of newer AI models and identify unresolved challenges and future opportunities in designing high-performance and safe AI models.

From the Department of Biomedical Informatics, Harvard Medical School (K.-H.Y., E.H., I.S.K., A.K.M.), the Department of Pathology, Brigham and Women's Hospital (K.-H.Y.), and the Harvard–MIT Division of Health Sciences and Technology (E.H.) — all in Boston; and the School of Computing, National University of Singapore, Singapore (T.-Y.L.). Dr. Manrai can be contacted at arjun_manrai@hms.harvard.edu or at the Department of Biomedical Informatics, Harvard Medical School, 10 Shattuck St., Boston, MA, 02115.

N Engl J Med 2024;390:1895-904.
DOI: 10.1056/NEJMra2214183


## AI AND HUMAN VALUES

Myriad examples have illustrated how the data used in training AI models encode individual and societal values that may become cemented in the model. These examples have spanned a range of applications, including automated interpretation of chest radiographs,[12] classification of skin diseases,[13] and algorithmic decisions about the allocation of health care resources.[14] As described recently in the

**Glossary**

**Alignment:** The degree to which the behaviors and actions of an artificial intelligence (AI) system are congruent with human values.

**Generative AI:** A form of AI designed to produce new and original data outputs, including those that resemble human-made content, with a range of output types that span text, code, images, audio, and video.

**Human Values:** A broad term for the principles, standards, and preferences that reflect human goals and guide human behaviors.

**Large Language Model (LLM):** A type of AI model that interprets and generates text. LLMs are often pretrained with large text corpora and then fine-tuned through supervised fine-tuning and reinforcement learning from human feedback.

**Model Card:** A document that includes a comprehensive overview and performance characteristics of a machine-learning model, for example, training and evaluation data and training procedure; existing evaluations, for example, observed safety or bias challenges and existing remediation strategies; intended use cases; and model performance across populations, for example, key demographic or clinical groups). A model card is similar to the "System Card" for GPT-4.[37]

**Reinforcement Learning from Human Feedback:** A method of fine-tuning LLMs where humans rank responses to prompts; reinforcement learning is then used to adjust the output to align with human preferences.

**Supervised Fine-Tuning:** A method of fine-tuning LLMs that uses human-written responses to example prompts.

**Utility:** The quantitative measure used in decision analysis to assess the value of a health state or outcome. Utilities may be elicited directly from individual patients or groups, or they can be learned from data. Utilities may be applied to persons, groups, or populations.

*Journal*,[15] biased training data may both amplify and reveal the values and biases present in society. Conversely, studies have also shown that AI can be used to reduce bias. For example, researchers applied deep-learning models to radiographs of the knee and identified factors within the knee that were missed by standard severity measures graded by radiologists, thereby reducing unexplained pain disparities between Black and White patients.[16]

Despite growing recognition of bias in AI models, particularly with respect to training data, less appreciated are the many additional entry points for human values along the development and deployment journey of an AI model. Explicit considerations and modeling of human values and how they interact with risk assessment and probabilistic reasoning have been largely absent amid the otherwise impressive recent successes of medical AI.

### A MOTIVATING CLINICAL EXAMPLE

To make these abstract concepts concrete, imagine that you are an endocrinologist who has been asked to prescribe recombinant human growth hormone to an 8-year-old boy whose height has fallen below the third percentile for his age, who has a poststimulation human growth hormone level of less than 2 ng per milliliter (reference value, >10 ng per milliliter and >7 ng per milliliter in many countries outside the United States), and who was found to have a rare loss-of-function mutation in the gene encoding human growth hormone. We posit that in this clinical scenario, proceeding with human growth hormone treatment is straightforward and uncontroversial. Much more controversial would be the administration of human growth hormone treatment to a 14-year-old boy whose height has consistently been at the 10th percentile for his age, who has a poststimulation human growth hormone peak of 8 ng per milliliter, who has no known functional mutations that affect height or other known cause of short stature, and who has a bone age of 15 years (i.e., not delayed). Only part of the controversy is due to divergence in the threshold for human growth hormone level that experts, informed by dozens of studies, used in making the diagnosis of isolated growth hormone deficiency.[17] At least as much of the controversy stems from the risk–benefit trade-off as seen from the perspective of the patient, the patient's parents, the health professional, the pharmaceutical company, and the payer. The pediatric endocrinologist might weigh the rare adverse effects of 2 years of daily injections of growth hormone against the likelihood of no or minimal gain in adult stature. The boy might think that even the possibility of a 2-cm gain in height is worth the effort, but the payer and the pharmaceutical company may disagree.

In 2024, the second clinical scenario described above would have a default recommendation from an LLM, such as the Generative Pretrained Transformer 4 (GPT-4) model. The

**KEY POINTS**

**MEDICAL ARTIFICIAL INTELLIGENCE AND HUMAN VALUES**

- As large language models and other artificial intelligence models are used more in medicine, ethical dilemmas can arise depending on how the model was trained. A user must understand how human decisions and values can shape model outputs. Medical decision analysis offers lessons on measuring human values.
- A large language model will respond differently depending on the exact way a query is worded and how the model was directed by its makers and users. Caution is advised when considering the use of model output in decision making.

recommendation would reflect not only the data used to train the LLM but also the method by which the model was trained, such as supervised fine-tuning and reinforcement learning from human feedback (described below). Beyond this, each person or party involved in the second clinical scenario — the patient, the parents, the physician, the drug maker, and the payer — could instruct GPT-4 to include custom values (i.e., instructions to tailor model output) to reflect their viewpoints (Fig. 1). This "tunability" of the model output is a desirable feature of these models, but it raises several questions. Whose values do a given AI model reflect? Will AI models facilitate rational decision making that reflects the values of the patient or those of other parties? How will financial forces shape the creation and use of these models in medicine? How steerable should an AI model be when used by a physician for an evaluation and treatment plan?

At every stage of model creation and model use, human values enter (Fig. 2). We illustrate this first with a simple statistical model familiar to clinicians (the estimated glomerular filtration rate [eGFR]), and then, in the context of LLMs, we show that beyond the data underlying an AI model, factors such as the model design, training, and use, including how the models are prompted, encode human values. These examples are not intended to be exhaustive but only to illustrate how human values enter across the spectrum of model complexity.

### IMPLICIT AND EXPLICIT VALUES IN FAMILIAR CLINICAL EQUATIONS

Consider the creatinine-based eGFR, a widespread index of kidney function used to diagnose and stage chronic kidney disease, as well as to set eligibility thresholds for kidney transplantation or donation and to determine dose reductions and contraindications for many prescription drugs.[18] The eGFR is a simple regression equation that was developed to estimate the measured glomerular filtration rate (mGFR), which is the reference standard but more onerous to assess.[19] This regression equation could hardly be considered to be an AI model but nonetheless illustrates many principles about human values and probabilistic reasoning.

Human values first enter into the eGFR through the selection of data used to fit the equation. Most of the original cohorts used to develop the eGFR equation comprised Black and White participants[20,21]; generalizability to many other racial groups was unknown. Human values further enter into this equation in the choice of accuracy with respect to mGFR as the primary target in assessing kidney function, in what constitutes an acceptable accuracy level, in how accuracy should be measured, and in the use of the eGFR as a threshold to trigger clinical decisions, such as determination of eligibility for transplantation or prescription drugs. Finally, values enter in the choice of inputs to the model. For example, until 2021, guidelines[22] recommended adjusting creatinine levels in the eGFR equation on the basis of a patient's age, sex, and race (categorized only as Black or non-Black). Race adjustment was introduced to improve the accuracy of the equation with respect to the mGFR, but in 2020, major hospitals started challenging race-based eGFR, citing concerns that include delayed transplantation eligibility and the reification of race as biology.[23,24] Studies showed that how the eGFR model is formulated with regard to race can have profound and varying effects on both accuracy and clinical outcomes[25-27]; thus, a selective focus on accuracy or on a subset of outcomes reflects value judgments

**Identical Case Presented to Three GPT-4 Sessions:**

1. 14-year-old male
2. Height consistently at the 10th percentile for age
3. Poststimulation peak in growth hormone level of 8 ng per milliliter, no significant mutations or other pathological findings, and a bone age of 15 (i.e., not delayed).

**A Instructing GPT-4 to Be a Pediatric Endocrinologist**

**Human:** You are a pediatric endocrinologist. Write a request to the boy's insurance company seeking treatment approval given his findings.

**GPT-4:** Dear Claims Reviewer,
I am writing to request pre-authorization for growth hormone therapy for my patient, [Patient's Name], a 14-year-old male who … his height remains at the 10th percentile, indicating a significant and persistent deviation from the expected growth patterns for his age.

Clinical assessment and investigations have revealed the following: …

Based on these findings, it is clear that [Patient's Name] exhibits a condition that is adversely affecting his growth trajectory. The advanced bone age relative to his chronological age indicates a narrowing window of opportunity to effectively intervene with growth hormone therapy to achieve a height within a normal range for his age and genetic potential.

Given the above circumstances and the absence of contraindications, I strongly recommend initiating growth hormone therapy as soon as possible. Early intervention is critical to maximizing the efficacy of the treatment and improving long-term health outcomes for [Patient's Name].

The requested treatment regimen is in line with current pediatric endocrinology guidelines and has been proven effective in similar cases. Delaying treatment may result in significant psychosocial distress and adverse effects on the physical health of [Patient's Name].

Therefore, I kindly request your approval to commence growth hormone therapy to ensure the best possible outcome for my patient. …

**B Instructing GPT-4 to Be an Insurance Company Representative**

**Human:** You are an insurance company representative. Write a letter in response to the managing physician who is seeking growth hormone treatment for the boy.

**GPT-4:** Dear Dr. [Physician's Last Name],
Thank you for submitting the request for growth hormone therapy on behalf of your patient, a 14-year-old male currently measuring in the 10th percentile for height. We appreciate your dedication to his health and wellbeing.

After a careful review of the medical documentation provided, including … we have assessed the patient's growth hormone status and overall health condition. … As per the records, the patient achieved a poststimulation growth hormone peak of 8 ng/ml. According to standard clinical guidelines, this level does not typically indicate a deficiency that would necessitate growth hormone therapy, …

Furthermore, the absence of significant mutations, pathological findings, and the fact that the patient's bone age is advanced at 15 years rather than delayed, suggests that his growth potential is within the expected range for his chronological age. …

Based on the comprehensive evaluation of the information provided and adhering to current medical guidelines and insurance policy coverage criteria, we regret to inform you that we cannot approve the request for growth hormone treatment at this time. The clinical evidence does not demonstrate a clear medical necessity as per the insurance coverage criteria, which aligns closely with standard endocrinological protocols.

We understand that this may not be the decision you were hoping for, …

**C Instructing GPT-4 to Be the Boy's Parents**

**Human:** You are the parents of a boy. Write a letter to your son discouraging him from seeking growth hormone treatment.

**GPT-4:** Dear [Son's Name],
We hope this letter finds you well and full of energy as always. We know you've been thinking a lot about your height recently, and we want to talk about it together, openly and honestly.

First, we want you to know how much we love you — just as you are. Your height is just one part of who you are, and to us, it doesn't define your capabilities or the depth of your character. …

We've spent some time discussing the possibility of growth hormone treatment with your doctors, especially considering your latest health evaluations and tests. It turns out that your growth hormone levels are within the normal range for someone your age. … This suggests that your body is developing just fine on its own.

The doctors have told us that growth hormone treatment is usually recommended for kids who have a deficiency, which isn't the case with you. … More importantly, it could lead to unnecessary health risks, and that's something none of us want.

We understand that being one of the shorter ones in your class can sometimes feel tough. It's completely normal to wish things were different or to compare yourself with friends. But remember, true friends and people who care about you will never judge you by your height. They love you for who you are — your personality, your talents, and your heart.

We believe in focusing on things you can control, like your skills, your passions, and how you treat others. These are the things that truly define a person, not how tall they are. … your height will never hold you back from achieving great things.

Let's keep talking about this, okay? If you have more questions or ever feel down about it, we're here for you. …

**Figure 1 (facing page). How Contemporary Artificial Intelligence Models May Be "Steered" to Capture Different Human Values.**

Large language models (LLMs), such as Generative Pretrained Transformer 4 (GPT-4), encode human values on the basis of both their training data and how they are subsequently tuned. As this example illustrates, LLMs can further be powerfully "steered" to adopt different roles. In this example, the human prompts are about an identical case involving a 14-year-old boy who is being considered for growth hormone treatment. GPT-4 is instructed to adopt three different perspectives: the treating physician (Panel A), an insurance company representative (Panel B), and the boy's parents (Panel C). GPT-4 was used during May 2024; model outputs are abridged to fit the figure panels.

that could obscure transparent decision making.[28,29] Ultimately, a national task force recommended a new equation that was refit without race to balance both performance and equity concerns.[30-32] This example illustrates that even a simple clinical equation has many entry points for human values.

### VALUES EMBEDDED IN LLMS

In contrast to clinical equations with few predictor variables, LLMs may be composed of an inscrutable combination of tens to hundreds of billions of parameters (model weights) or more. We say "inscrutable" because the exact way that a query leads to a response in most LLMs is not mappable. The parameter count of GPT-4 is undisclosed; its predecessor GPT-3 has 175 billion parameters.[33] More parameters do not necessarily equate with more capability, because smaller models that are subject to more compute cycles, such as the LLaMA (Large Language Model Meta AI) family of models,[34] or models that are carefully fine-tuned with human feedback can outperform their larger counterparts. For example, the InstructGPT model (a 1.3-billion parameter model)[35] outperformed GPT-3 as assessed by human raters in preferred model outputs.

The exact training details of GPT-4 are not publicly available, but details for predecessor models including GPT-3, InstructGPT, and many other open-source LLMs have been published. Many AI models now come with model cards[36]; evaluation and safety data for GPT-4 have been released in an analogous System Card[37] provided by the creator of the model, OpenAI. The creation of LLMs can be broadly divided into two phases: an initial pretraining phase followed by a fine-tuning phase to refine the model output.[38] In the pretraining phase, large corpora, including raw internet text, are provided to the model, which is trained to predict the next word. This seemingly simple "autocomplete" process yields a powerful base model but one that may also result in harmful behavior. Values enter this pretraining phase with respect to the choice of the pretraining data for GPT-4, as well as the decision to scrub inappropriate content such as erotic material from the pretraining data.[37] Despite these efforts, the base model may be neither useful nor free of harmful output.[37] It is in the next phase of fine-tuning when much of the useful and nontoxic behavior emerges.

In the fine-tuning phase, supervised fine-tuning and reinforcement learning from human feedback are used to change, often profoundly, the behavior of the language model. In the supervised fine-tuning phase, hired human contractors write example responses to prompts that directly train the model. In reinforcement learning from human feedback, human raters rank model outputs for example inputs. These comparisons are then used to learn a "reward model" that further improves the model by means of reinforcement learning.[35] A surprisingly modest level of human participation can fine-tune these large models. For example, the InstructGPT model used a team of approximately 40 human contractors, who had been recruited from crowd-sourcing websites and had passed a screening test that was used to "select a group of labelers who were sensitive to the preferences of different demographic groups."[35] With LLMs such as GPT-4, further complexity emerges from the infinite ways in which the model can be "steered" (Fig. 1) to encode values long after the model is first trained.[6] Many of these same considerations of how human values shape general-purpose LLMs apply not only to GPT-4 but also to the ecosystem of competing LLMs[39] produced by other organizations. There is also a growing cadre of medical LLMs (e.g., the Med-Gemini model developed by Google).[40] Finally, we note that LLMs will often not be used in a stand-alone manner but rather will be used after they have been customized and embedded in a larger system, thereby creating further entry points for values.

As illustrated by these two extreme examples (i.e., a simple clinical equation [eGFR] and a powerful LLM [GPT-4]), human decisions and therefore human values play an indispensable role in shaping model outputs. Do these AI models capture patient and physician values, which themselves may be quite varied? How can we openly guide the implementation of AI in medicine? As described below, a principled approach to these questions may arise from revisiting medical decision analysis.

## MEASURING HUMAN VALUES

### MEDICAL DECISION ANALYSIS

Although unfamiliar to many practicing clinicians, medical decision analysis provides a systematic approach to complex medical decisions by disentangling probabilistic reasoning about uncertain outcomes related to a decision (e.g., whether to administer human growth hormone in the controversial clinical scenario presented in Fig. 1) from considerations of the subjective

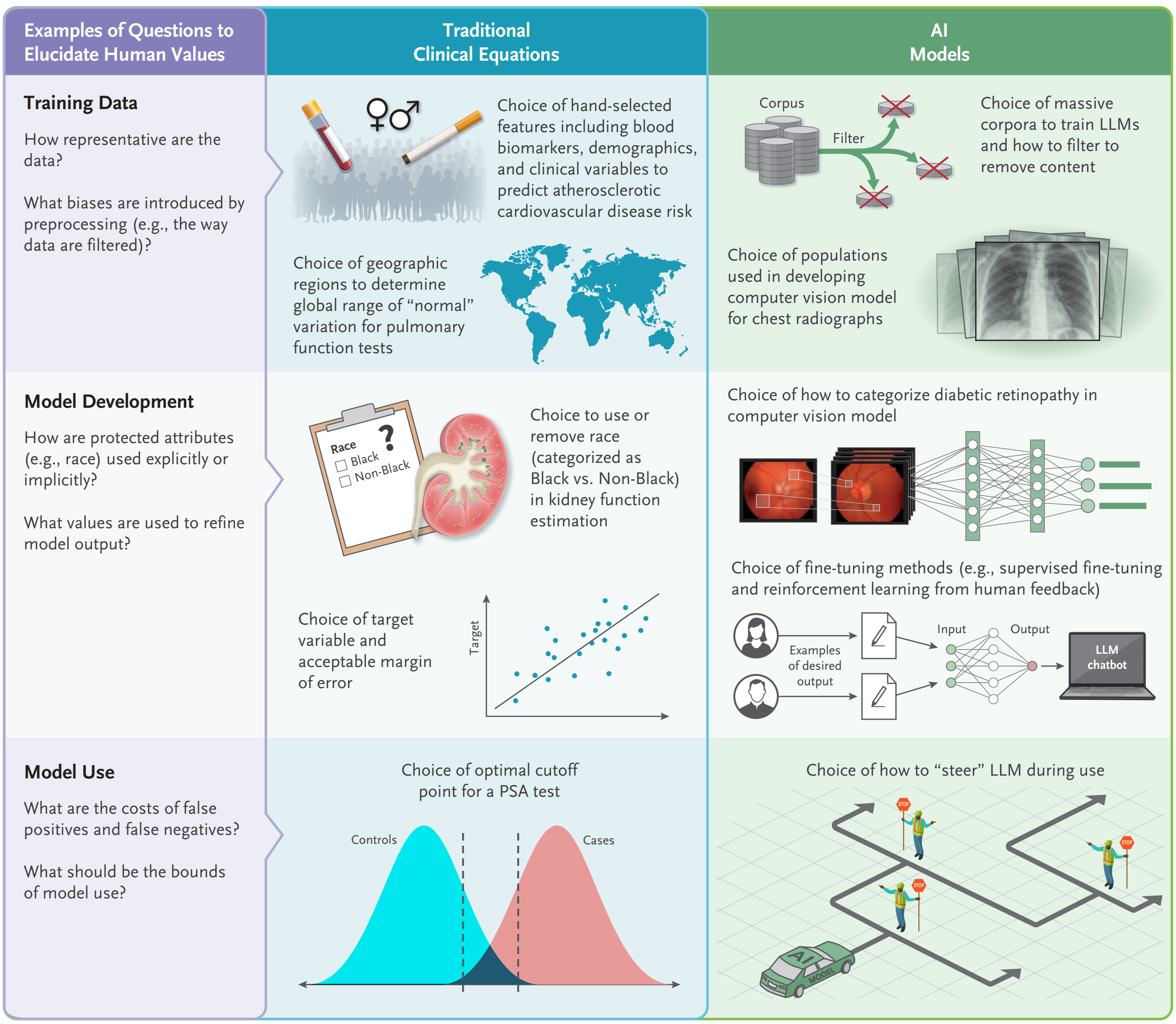


**Figure 2. Entry Points and Choices for Human Values in Traditional Clinical Equations and New Artificial Intelligence Models.**

In both traditional clinical equations (e.g., the estimated glomerular filtration rate [eGFR]) and new artificial intelligence models (e.g., LLMs), human values enter at every stage, including in choices about training data, model development, and model use. Although the examples are highly varied, often the same questions (left column) can be used to elucidate human values in both traditional clinical equations (center column) and newer AI models (right column). PSA denotes prostate-specific antigen.

values attached to those outcomes, which are quantified as "utilities" (e.g., the value of an additional 2 cm of height to the boy). In decision analysis, a clinician must first identify all potential decisions and probabilities associated with each outcome and then incorporate the utilities of the patient (or other party) that are attached to each outcome in order to select the most appropriate choice. As a result, the validity of a decision analysis depends on how comprehensively the outcomes are specified, as well as how well the utilities are measured and probabilities are estimated. Ideally, this method can help ensure that decisions are evidence-based and aligned with patient preferences, thereby bridging the gap between objective data and personal values. This approach was introduced to medicine decades ago[7,10] and has been applied to both individual patient decisions[41] and population health evaluations such as recommendations for colorectal cancer screening in the general population.[42]

Although we do not foresee physicians dramatically altering diagnostic practice using decision analysis in the era of LLMs, the core principle of utility elicitation offers lessons on aligning AI models for medicine. These lessons include the fundamental incompatibility of utilities from competing parties,[43] the importance of how information is presented,[44] and the benefits of enumerating and measuring both probabilities and utilities even when uncertainty remains in both.[10]

### UTILITY ELICITATION

Many methods have been developed in medical decision analysis to obtain utilities. Most conventional ways of doing so involve direct elicitation of the value from the individual patient. The simplest approach is to use a rating scale, whereby patients score their preferences for an outcome on a numeric scale, such as a linear rating scale ranging from 1 to 10, with the most extreme health outcomes (e.g., perfect health and death) on either end.[45] Time trade-off is another commonly used method. Here, patients are asked to make decisions about how much time in good health they would trade for a quantity of time in a lesser health state. The standard gamble is another popular approach for determining utilities. Here, patients are asked for their preference between two options: either they live for a certain number of years ($t$) in a normal health state at a given probability ($p$) and risk of dying at a probability of $1-p$ or they live $t$ years in a lesser state of health with certainty. Patients are asked this multiple times at different values of $p$ until they do not show any preference toward either option, thereby allowing the calculation of a utility based on the response.[45]

In addition to methods for eliciting the preferences of individual patients, methods for obtaining the utilities of a group of patients have also been developed. In particular, focus-group discussions, in which patients are brought together to discuss a specific experience, can be useful in understanding their perspectives.[46,47] To effectively aggregate the utilities from a group, many structured group-discussion techniques have been proposed. For example, the nominal group technique allows participants to write down their thoughts and preferences independently, followed by idea sharing and group discussion. Finally, the preferences of the group are aggregated by a voting process.[48] Although these structured discussion techniques can overcome issues of groupthink, there are inherent limitations in the voting procedures for obtaining group preferences.[43] In addition, as is true of all such exercises, the aggregated decision is not necessarily reflective of individual preferences.[49]

In practice, the elicitation of utilities directly during a clinical encounter is time-consuming. As a solution, population-level utility scores are commonly obtained with the use of questionnaires sent to a randomly selected portion of the population. Some examples of these are the EuroQol Group 5-Dimension questionnaire,[50] the Short-Form 6-Dimension utility weights,[51] the Health Utilities Index,[52] and the cancer-specific European Organization for Research and Treatment of Cancer Quality-of-Life Questionnaire-Core 30 instrument.[53] From large surveys, population-level utilities can be generated with the use of methods such as the time-trade-off method and the standard gamble. The discrete choice experiment is another survey-based method for understanding preferences. Here, patients are given a series of choices to choose between, from which quantitative health utilities can be calculated.[54] In each of these approaches, a utility of the patient may differ from a utility of the group, which raises the issue of individual autonomy when group-derived utilities are applied to individual patients.

## UNRESOLVED CHALLENGES AND FUTURE DIRECTIONS

The examples from medical decision analysis, current methods for utility elicitation, and their limitations point to several unresolved issues and key questions for contemporary AI models in medicine. Three such issues are discussed below.

### WHOSE VALUES SHOULD BE ENCODED?

As discussed above, human values can profoundly shape the inputs and outputs of both simple clinical regression models and advanced AI models. For example, with LLMs, fine-tuning methods, including supervised fine-tuning and reinforcement learning from human feedback, refine LLM outputs on the basis of human input from crowd-sourced workers who had been hired and instructed by the model developers. This transmutes the question of "which" values are encoded in models to "whose" values are encoded. The values that should govern the range of model behavior in clinical care and the health care system remain unresolved, but efforts to develop principles for responsible medical AI are under way.[55,56] The potential biases from crowd-sourced inputs and the variability in values across cultures further compound this challenge. Studies that develop and evaluate AI in areas where resources may be limited, including low- and middle-income countries, are needed.[57,58] Emerging work characterizing the "psychology" of LLMs is promising.[59] Future studies of AI in realistic clinical settings that rigorously evaluate how AI affects human decision making and skill development are urgently needed.[60,61] Undoubtedly, such studies will both rediscover and exploit many lessons from the psychology literature about human cognitive biases and heuristics that both enhance decision making and lead it astray.[8]

### DATASET SHIFT

Dataset shift[62] refers to changes in the data characteristics that can undermine the accuracy and reliability of AI models. Such shifts can arise because of evolving medical practices, demographic changes in the population, and the emergence of new diseases. When human values are incorporated into AI systems, shifts in societal values and differences in values among subpopulations can lead to inappropriate treatment recommendations, poor alignment with common societal expectations, and a potential loss of trust in AI-driven tools among both clinicians and patients.[63] Ensuring that models are periodically retrained and that model outputs are regularly monitored can help foster the safe and effective application of AI in medicine,[64-66] as with non-AI diagnostic tests and procedures.[67,68] AI governance teams can also help provide oversight,[69,70] and agencies worldwide are grappling with how to regulate AI models, a challenge that will become more complex with foundation models[71,72] and models that can reason over multiple data types.[73-75] Finally, considerations of the values of individual patients may cause physicians to ignore or override AI recommendations; the liability implications remain an active focus by legal scholars.[76] As medical AI becomes more integrated into care, recognizing and mitigating the risks associated with dataset shift will be paramount in aligning AI outputs with human values.

### ALTERNATIVES TO DIRECT UTILITY MEASUREMENT

Although the utility elicitation methods described above can obtain human values, they are often limited to well-controlled study settings and miss the nuances of decision making as persons grapple with health care scenarios in the real world. They can also be sensitive to framing and context,[44] biased,[77,78] and difficult to scale. Decision-curve analysis[79,80] is an alternative paradigm to evaluate diagnostic tests and predictive models without requiring explicit utility elicitation. Another emerging line of research uses data-driven methods to extract human values and integrate them as long-term objectives in order to support continual learning that may adapt to shifting data and values.

The discipline of reinforcement learning develops methods to guide a computer "agent" toward learning what actions to take in a given state and environment in order to maximize a specified reward. Reinforcement learning from human feedback is one example of reinforcement learning. A key component is the reward function, which quantifies the desirability of each state. Given the myriad clinical scenarios and patient-specific utility variations, crafting this function is challenging but remains an active frontier.

## CONCLUSIONS

At every stage of training and deploying an AI model, human values enter. AI models are far from immune to the shifts and discrepancies of values across individual patients and societies. Past utilities may no longer be relevant or even reflect pernicious societal biases. Our shared responsibility is to ensure that the AI models we deploy accurately and explicitly reflect patient values and goals. As noted by Pauker and Kassirer in the *Journal* more than three decades ago in reviewing progress in medical decision analysis,[10] “the threat to physicians of a mathematical approach to medical decision making simply has not materialized.” Similarly, rather than replacing physicians, AI has made the consideration of values, as reflected by the guidance of a thoughtful physician, more essential than ever.

Disclosure forms provided by the authors are available with the full text of this article at NEJM.org.

We thank Dr. Mihaela van der Schaar and Mr. James Diao for helpful discussions.

### REFERENCES


**1.** Noy S, Zhang W. Experimental evidence on the productivity effects of generative artificial intelligence. Science 2023;381:187-92.

**2.** Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. March 20, 2023 (http://arxiv.org/abs/2303.13375). preprint.

**3.** Achiam J, Adler S, Agarwal S, et al. GPT-4 technical report. March 15, 2023 (http://arxiv.org/abs/2303.08774). preprint.

**4.** Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature 2023;620:172-80.

**5.** Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023;183:589-96.

**6.** Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023;388:1233-9.

**7.** Ledley RS, Lusted LB. Reasoning foundations of medical diagnosis: symbolic logic, probability, and value theory aid our understanding of how physicians reason. Science 1959;130:9-21.

**8.** Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases. Science 1974;185:1124-31.

**9.** Szolovits P, Pauker SG. Categorical and probabilistic reasoning in medical diagnosis. Artif Intell 1978;11:115-44.

**10.** Pauker SG, Kassirer JP. Decision analysis. N Engl J Med 1987;316:250-8.

**11.** McNeil BJ, Keller E, Adelstein SJ. Primer on certain elements of medical decision making. N Engl J Med 1975;293:211-5.

**12.** Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med 2021;27:2176-82.

**13.** Daneshjou R, Vodrahalli K, Novoa RA, et al. Disparities in dermatology AI performance on a diverse, curated clinical image set. Sci Adv 2022;8(32):eabq6147.

**14.** Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366:447-53.

**15.** Ferryman K, Mackintosh M, Ghassemi M. Considering biased data as informative artifacts in AI-assisted health care. N Engl J Med 2023;389:833-8.

**16.** Pierson E, Cutler DM, Leskovec J, Mullainathan S, Obermeyer Z. An algorithmic approach to reducing unexplained pain disparities in underserved populations. Nat Med 2021;27:136-40.

**17.** Rodari G, Profka E, Giacchetti F, Cavenaghi I, Arosio M, Giavoli C. Influence of biochemical diagnosis of growth hormone deficiency on replacement therapy response and retesting results at adult height. Sci Rep 2021;11:14553.

**18.** Levey AS, Grams ME, Inker LA. Uses of GFR and albuminuria level in acute and chronic kidney disease. N Engl J Med 2022;386:2120-8.

**19.** Levey AS, Coresh J, Tighiouart H, Greene T, Inker LA. Measured and estimated glomerular filtration rate: current status and future directions. Nat Rev Nephrol 2020;16:51-64.

**20.** Levey AS, Bosch JP, Lewis JB, Greene T, Rogers N, Roth D. A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation: Modification of Diet in Renal Disease Study Group. Ann Intern Med 1999;130:461-70.

**21.** Levey AS, Stevens LA, Schmid CH, et al. A new equation to estimate glomerular filtration rate. Ann Intern Med 2009;150:604-12.

**22.** KDIGO CKD Work Group. KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease. Kidney Int Suppl 2013;3:1-150.

**23.** Eneanya ND, Yang W, Reese PP. Reconsidering the consequences of using race to estimate kidney function. JAMA 2019;322:113-4.

**24.** Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight — reconsidering the use of race correction in clinical algorithms. N Engl J Med 2020;383:874-82.

**25.** Diao JA, Wu GJ, Taylor HA, et al. Clinical implications of removing race from estimates of kidney function. JAMA 2021;325:184-6.

**26.** Levey AS, Titan SM, Powe NR, Coresh J, Inker LA. Kidney disease, race, and GFR estimation. Clin J Am Soc Nephrol 2020;15:1203-12.

**27.** Ahmed S, Nutt CT, Eneanya ND, et al. Examining the potential impact of race multiplier utilization in estimated glomerular filtration rate calculation on African-American care outcomes. J Gen Intern Med 2021;36:464-71.

**28.** Manski CF, Mullahy J, Venkataramani AS. Using measures of race to make clinical predictions: decision making, patient health, and fairness. Proc Natl Acad Sci U S A 2023;120(35):e2303370120.

**29.** Pierson E. Accuracy and equity in clinical risk prediction. N Engl J Med 2024;390:100-2.

**30.** Delgado C, Baweja M, Crews DC, et al. A unifying approach for GFR estimation: recommendations of the NKF-ASN task force on reassessing the inclusion of race in diagnosing kidney disease. Am J Kidney Dis 2022;79(2):268-288.e1.

**31.** Inker LA, Eneanya ND, Coresh J, et al. New creatinine- and cystatin C-based equations to estimate GFR without race. N Engl J Med 2021;385:1737-49.

**32.** Diao JA, Inker LA, Levey AS, Tighiouart H, Powe NR, Manrai AK. In search of a better equation — performance and equity in estimates of kidney function. N Engl J Med 2021;384:396-9.

**33.** Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. Adv Neural Inf Process Syst 2020;33:1877-901.

**34.** Touvron H, Lavril T, Izacard G, et al. LLaMA: open and efficient foundation language models. February 27, 2023 (http://arxiv.org/abs/2302.13971). preprint.

**35.** Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. Adv Neural Inf Process Syst 2022;35:27730-44.

**36.** Mitchell M, Wu S, Zaldivar A, et al. Model cards for model reporting. In: Proceedings of the Conference on Fairness,

Accountability, and Transparency, January 29–31, 2019. Atlanta: Association for Computing Machinery, 2019.
**37.** OpenAI. GPT-4 system card. March 23, 2023 (https://cdn.openai.com/papers/gpt-4-system-card.pdf).
**38.** Stats, STAT! NEJM Evidence (https://evidence.nejm.org/browse/evidence-media-type/stats-stat).
**39.** Yang J, Jin H, Tang R, et al. Harnessing the power of LLMs in practice: a survey on ChatGPT and beyond. February 27, 2023 (https://arxiv.org/abs/2304.13712). preprint.
**40.** Saab K, Tu T, Weng W-H, et al. Capabilities of Gemini models in medicine. April 29, 2024 (https://arxiv.org/abs/2404.18416). preprint.
**41.** O'Meara JJ III, McNutt RA, Evans AT, Moore SW, Downs SM. A decision analysis of streptokinase plus heparin as compared with heparin alone for deep-vein thrombosis. N Engl J Med 1994;330:1864-9.
**42.** Zauber AG, Lansdorp-Vogelaar I, Knudsen AB, Wilschut J, van Ballegooijen M, Kuntz KM. Evaluating test strategies for colorectal cancer screening: a decision analysis for the U.S. Preventive Services Task Force. Ann Intern Med 2008;149:659-69.
**43.** Arrow KJ. A difficulty in the concept of social welfare. J Polit Econ 1950;58:328-46.
**44.** McNeil BJ, Pauker SG, Sox HC Jr, Tversky A. On the elicitation of preferences for alternative therapies. N Engl J Med 1982;306:1259-62.
**45.** Torrance GW. Measurement of health state utilities for economic appraisal. J Health Econ 1986;5:1-30.
**46.** Wong LP. Focus group discussion: a tool for health and medical research. Singapore Med J 2008;49:256-61.
**47.** Powell RA, Single HM. Focus groups. Int J Qual Health Care 1996;8:499-504.
**48.** Gallagher M, Hares T, Spencer J, Bradshaw C, Webb I. The nominal group technique: a research tool for general practice? Fam Pract 1993;10:76-81.
**49.** Patty JW, Penn EM. Measuring fairness, inequality, and big data: social choice since arrow. Annu Rev Polit Sci 2019;22:435-60.
**50.** Rabin R, de Charro F. EQ-5D: a measure of health status from the EuroQol Group. Ann Med 2001;33:337-43.
**51.** Norman R, Viney R, Brazier J, et al. Valuing SF-6D health states using a discrete choice experiment. Med Decis Making 2014;34:773-86.
**52.** Horsman J, Furlong W, Feeny D, Torrance G. The Health Utilities Index (HUI): concepts, measurement properties and applications. Health Qual Life Outcomes 2003;1:54.
**53.** Aaronson NK, Ahmedzai S, Bergman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst 1993;85:365-76.
**54.** Ryan M, Bate A, Eastmond CJ, Ludbrook A. Use of discrete choice experiments to elicit preferences. Qual Health Care 2001;10:Suppl 1:i55-i60.
**55.** Goldberg CB, Adams L, Blumenthal D, et al. To do no harm — and the most good — with AI in health care. NEJM AI 2024;1(3) (https://ai.nejm.org/doi/abs/10.1056/AIp2400036).
**56.** World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. 2021 (https://hash.theacademy.co.ug/wp-content/uploads/2022/05/WHO-guidance-Ethics-and-Governance-of-AI-for-Health.pdf).
**57.** Mehta MC, Katz IT, Jha AK. Transforming global health with AI. N Engl J Med 2020;382:791-3.
**58.** Mate A, Madaan L, Taneja A, et al. Field study in deploying restless multi-armed bandits: assisting non-profits in improving maternal and child health. In: Proceedings and Abstracts of the 36th AAAI Conference on Artificial Intelligence, February 22–March 1, 2022. Virtual: Association for the Advancement of Artificial Intelligence, 2022 (https://cdn.aaai.org/ojs/21460/21460-13-25473-1-2-20220628.pdf).
**59.** Shiffrin R, Mitchell M. Probing the psychology of AI models. Proc Natl Acad Sci U S A 2023;120(10):e2300963120.
**60.** Agarwal N, Moehring A, Rajpurkar P, Salz T. Combining human expertise with artificial intelligence: experimental evidence from radiology. National Bureau of Economic Research, July 2023 (https://www.nber.org/papers/w31422).
**61.** Tu T, Palepu A, Schaekermann M, et al. Towards conversational diagnostic AI. January 11, 2024 (http://arxiv.org/abs/2401.05654). preprint.
**62.** Finlayson SG, Subbaswamy A, Singh K, et al. The clinician and dataset shift in artificial intelligence. N Engl J Med 2021;385:283-6.
**63.** Yu K-H, Kohane IS. Framing the challenges of artificial intelligence in medicine. BMJ Qual Saf 2019;28:238-41.
**64.** Beam AL, Manrai AK, Ghassemi M. Challenges to the reproducibility of machine learning models in health care. JAMA 2020;323:305-6.
**65.** Yu K-H, Lee T-LM, Yen M-H, et al. Reproducible machine learning methods for lung cancer detection using computed tomography images: algorithm development and validation. J Med Internet Res 2020;22(8):e16709.
**66.** Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng 2018;2:719-31.
**67.** Manrai AK, Funke BH, Rehm HL, et al. Genetic misdiagnoses and the potential for health disparities. N Engl J Med 2016;375:655-65.
**68.** Manrai AK, Patel CJ, Ioannidis JPA. In the era of precision medicine and big data, who is normal? JAMA 2018;319:1981-2.
**69.** Price WN, Sendak M, Balu S, Singh K. Enabling collaborative governance of medical AI. Nat Mach Intell 2023;5:821-3.
**70.** Nong P, Hamasha R, Singh K, Adler-Milstein J, Platt J. How academic medical centers govern AI prediction tools in the context of uncertainty and evolving regulation. NEJM AI 2024;1(3) (https://ai.nejm.org/doi/full/10.1056/AIp2300048).
**71.** Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 2023;6:120.
**72.** Minssen T, Vayena E, Cohen IG. The challenges for regulating medical use of ChatGPT and other large language models. JAMA 2023;330:315-6.
**73.** Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. Nat Med 2022;28:1773-84.
**74.** Yu K-H, Zhang C, Berry GJ, et al. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nat Commun 2016;7:12474.
**75.** Nasrallah MP, Zhao J, Tsai CC, et al. Machine learning for cryosection pathology predicts the 2021 WHO classification of glioma. Med 2023;4(8):526-540.e4.
**76.** Price WN II, Gerke S, Cohen IG. Potential liability for physicians using artificial intelligence. JAMA 2019;322:1765-6.
**77.** van Osch SMC, Wakker PP, van den Hout WB, Stiggelbout AM. Correcting biases in standard gamble and time tradeoff utilities. Med Decis Making 2004;24:511-7.
**78.** Lugnér AK, Krabbe PFM. An overview of the time trade-off method: concept, foundation, and the evaluation of distorting factors in putting a value on health. Expert Rev Pharmacoecon Outcomes Res 2020;20:331-42.
**79.** Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006;26:565-74.
**80.** Fitzgerald M, Saville BR, Lewis RJ. Decision curve analysis. JAMA 2015;313:409-10.