ICU
MANAGEMENT & PRACTICE

Artificial Intelligence and Human Values

Recent advancements in generative AI have produced large language models (LLMs) capable of writing persuasive articles, passing professional exams, and crafting empathetic messages. While their potential in medicine and healthcare is significant, concerns about their accuracy, reliability, and alignment with human values persist. These concerns include risks of confabulation and factual inaccuracies. As these risks are addressed, questions arise about the human values embedded in AI models and how they may diverge from human values, even without overt errors or toxic content.

Human values are integral to AI models, influencing their creation and implementation. This continues a historical trend in medical decision-making, where probabilistic models have long required value judgments. Early pioneers of medical decision analysis and subsequent scholars have worked to separate probabilities (chances of events) from utilities (quantified value judgments). The nuanced understanding of individual values and risks underscores the indispensable role of thoughtful clinicians.

Despite growing awareness of bias in AI models, the influence of human values throughout AI development and deployment is less recognised. Medical AI advancements largely lack explicit consideration of human values and their interaction with risk assessment and probabilistic reasoning.

Unlike clinical equations with few predictor variables, LLMs like GPT-4 consist of tens to hundreds of billions of parameters, making their inner workings complex and not easily interpretable. For example, GPT-3 has 175 billion parameters, but smaller models with more compute cycles or fine-tuned with human feedback can perform better.

LLMs are developed in two main phases: pretraining and fine-tuning. In the pretraining phase, models are trained on large text corpora to predict the next word in a sequence. This phase involves selecting appropriate pretraining data and removing harmful content, though the resulting base model may still exhibit undesirable behaviour.

The fine-tuning phase involves supervised fine-tuning and reinforcement learning from human feedback, which significantly refines the model's behaviour. Human contractors write example responses and rank model outputs to create a reward model, improving the LLM through reinforcement learning.

Human values enter LLMs during data selection and training and through steering the model after initial training. These values influence model outputs and can vary widely, raising questions about whose values the AI reflects and how AI implementation in medicine can be guided. This discussion is relevant not only for GPT-4 but also for other LLMs and medical-specific LLMs like Med-Gemini. The interplay of human values in model creation and deployment underscores the need for a principled approach to integrating AI in medicine.

Developing principles for responsible medical AI is ongoing, considering potential biases from crowd-sourced inputs and cultural variability in values. Studies in low- and middle-income countries are necessary, and future research should evaluate AI's impact on human decision-making and skill development in clinical settings.

Regulatory agencies worldwide are addressing the regulation of AI models, particularly foundation models and those reasoning over multiple data types. Considering individual patient values may cause physicians to override AI recommendations, raising liability issues.

The collective responsibility is to ensure that AI models accurately represent patient values and goals. AI does not replace physicians but emphasises the importance of incorporating values guided by thoughtful physicians into decision-making processes.

Source: NEJM
Image Credit: iStock

Published on : Tue, 4 Jun 2024