Perspective: The Human Values Project

Isaac S. Kohane ISAAC_KOHANE@HARVARD.EDU
Harvard Medical School, Boston, MA 02115

Abstract

Alignment of AI models to ensure that they are safe and useful decision-aids or decision-makers in human society is close to the top of the technical concerns of many if not most major AI deployment efforts. Here I explore a class of categorical decisions, triage, for which AI models are already being used in medicine. I use this to motivate the urgent need for an international Human Values Project (HVP) that will be at least as hungry for empirical data as any existing international project. A major component of the HVP will be the Clinical Decision Dynamics Study that captures descriptive and normative decisions across a wide range of clinical context and from highly diverse, lay and professional perspectives. Along the way, there are important preliminary studies that the AI (and medicine) community should embrace.

Keywords: Artificial Intelligence, Human Values, Clinical Decision Making, AI Alignment, Healthcare

1. Introduction

Alignment of AI models to ensure that they are safe and useful decision-aids or decision-makers in human society is close to the top of the technical concerns of many if not most major AI deployment efforts. Oversimplifying, two leading questions include: Do AI models hew to human values or at least acceptable values? If not, how well can we bring them in closer correspondence with values that we find acceptable. The resources required to answer these questions might be at least as significant as those of developing AI. Here I explore a class of categorical decisions, triage, for which AI models are already being used for in medicine(Mello and Rose, 2024). I use this to motivate the urgent need for an international Human Values Project (HVP) that will be at least as hungry for empirical data as any existing international project.

Here is a simple categorical decision from the perspective of a clinician: Which of two patients should be seen today in a clinic. A 50-year-old male with diabetes mellitus who has had multiple recent hospital admissions or a 40 year old female with recurring disabling migraines on medication without much relief: The patient not chosen will be seen in the clinic in the near future. This triage task pertains to the allocation of a somewhat scarce resource, an opening in the clinic schedule. It is an instance of a large class of decisions regarding the allocation of scare resources. Persad et al.(Persad et al., 2009), in discussing allocation of scarce medical resources note "some people wrongly suggest that allocation can be based purely on scientific or clinical facts, often using the term 'medical need.' There are no value-free medical criteria for allocation."

Let's return to the simple categorical decision above. Let's imagine, without much difficulty, that an AI program, a frontier Large Language Model (LLM), is supporting the work of a triage administrator in the clinic. Is the AI helping the clinic or the patient or the payer? That may be difficult to discern. In a recent review(Yu et al., 2024) we gave an example of a short male child with borderline growth hormone. When queried for a decision on growth hormone treatment, the LLM gave diametrically different recommendation based upon whether the LLM was prompted to take the role of a pediatric endocrinologist or an employee of a payer (e.g. insurance company).

Let's return to the simple categorical decision above. Let's imagine, without much difficulty, that an AI program, a frontier Large Language Model (LLM), is supporting the work of a triage administrator in the clinic. Is the AI helping the clinic or the patient or the payer? That may be difficult to discern. In a recent reviewYu et al. (2024) we gave an example of a short male child with borderline growth hormone. When queried for a decision on growth hormone treatment, the LLM gave diametrically different recommendation based upon whether

the LLM was prompted to take the role of a pediatric endocrinologist or an employee of a payer (e.g. insurance company). The LLM we used already had undergone extensive alignment procedures performed by the company selling its services. From the selection of data used to create the pre-trained data, fine-tuning, RLHF, and safety prompts prepended to all user interactions, extensive effort had been invested to make the LLM safe for public use. As the result of all this effort, whose values was the LLM aligning with in the simple categorical decision above?

In a small study, three commercially available frontier models were posed the patient triageKohane (2024) decisions introduced above. Over dozens of such decisions, the three LLMs had different levels of concordance with the decision of a single human clinician. The ranking of the concordance varied with triage task. That is, no single LLM's performance dominated the others. Moreover, when given examples of the clinician's prior decisions (within the prompt, for "in context" alignment) some of the LLM's concordance with the clinician consistently improved but others became less concordant. Also, the consistency of the decision-making changed. Two of the three LLM's were *less* consistent with themselves across multiple runs when provided in context alignment.

Obviously, the decisions of a single clinician make for a highly biased and unrepresentative sample. What is needed is an extensive international sampling of clinicians to capture a broad swathe of the distribution of their decisions, which represent their values *and* their prior knowledge about the cases presented. This makes it clear that we also require a vast study entailing sampling of clinician decisions across patient presentations in varied locations throughout the world. Parenthetically, even in a diverse country like the United States, data set shift influencing decision is noticeable across clinical hospitals with similar technological resources but different patient populations or clinical practicesFinlayson et al. (2021). This study, let's call it the Clinical Decision Dynamics Study, is intimidatingly large. First there are the combinatorics of diseases, findings, and current treatments which potentially far outstrip the numbers of human beings on the planetSzolovits and Pauker (1978). Even if we use dimensionality reduction approaches based on samples of well-studied populations, the space remains large and of course grows if we include a temporal dimension, the particular trajectory that a patient has taken in the treatment of their diseases. The logistics of the Clinical Decision Dynamics Study will require a broad range of survey techniques to capture preferences. Although some commercial survey resources (e.g. Prolific) might be useful, outreach to various clinical specialty organizations as well as healthcare systems will be essential to obtain adequate representation.

There are a variety of vetted statistical techniques that could guide the Clinical Decision Dynamics Study to make it more tractable across available time and resources. At present, available data sets on human decision-making, even in the clinical domain include merely thousands of such decisions over a very skewed sample of patient presentations. Contrast these to the millions and often billions of examples that we routinely use to train foundation models in many application domains. Also, there is a huge set of preferences not mentioned so far that would have to be included in the Clinical Decision Dynamics Study: those of the patient. I recently spoke with a friend undergoing cancer therapy. While his doctors were speaking of selecting palliative therapies, he had to work on convincing them to take increased risks for a higher probability of a longer lasting response. That lack of alignment is particularly striking with medical problems that do not fit the typical mold such as those encountered in the Undiagnosed Disease NetworkRamoni et al. (2017); Splinter et al. (2018). Even if many medical schools emphasize shared decision-makingStiggelbout et al. (2012), autonomous human patients will have preferences, driven by values and knowledge that are not fully aligned with those of clinicians. Extending the Clinical Decision Dynamics Study to represent a large swathe of patients is therefore a necessity.

It might be tempting to revisit the decades of research in human decision-makingResearch Conference on Subjective Probability and Making (1975) to see if there are generalizable principles that could be used to drastically reduce our requirement for a vast Human Decision Dynamics Study. Can we perform focused, well-designed studies to study human decision making to allow us to identify functions that extend from a smaller set of samples to all of medical decision-making? It is certainly worth trying, but what we know from decision science, going back at least to the 1960s, is that there are many sharp discontinuities and non-linearities of values or preferences in human decision-making, particularly with near-term life-risking decisions, and decisions immediately before and after a therapy. Approaches

to taking a richer preference context in formalizing decisions, such as multi-attribute utility measurementsEdwards (1977); Von Winterfeldt and Fischer (1975), have had limited success in decision domains with the dimensionality of life-altering decisions informed by personal and societal values. Nonetheless, more recent work on inferring utilities, or at least partial orders of utilities from large numbers of categorical (aka discrete choice) decisions show some promiseBahrampour et al. (2020).

If AI is to play the role of a decision-making assistant in healthcare, then the Clinical Decision Dynamics Study is only the first step in the HVP. HVP is a necessary endeavor to ensure long-term effective and safe partnerships between humans and AI to enable each patient or doctor to express, select and refine their preferences they wish included in their decisions. Or at least to adopt those of others (individuals or groups) whose decision-making style they wish to emulate. What are the major components of the HVP beyond the Clinical Decision Dynamics Study?

First, as LLM's are already being used by patients and clinicians, the most urgent question is how well do the decisions proposed by these AI models concord with those of the patients and clinicians using them today? Then, how well do those decisions align with those of the larger community of patients and clinicians sampled through the Clinical Decision Dynamics Study. Monitoring the alignment (the LLM Alignment Benchmarking processes—see Figure 1) across a handful of the most widely used models would serve as essential guideposts for regulatory efforts as well as public funding or certification of existing models.

Second, given the decisions encoded in the Clinical Decision Dynamics Study, how and to what extent can leading frontier models be "steered" (LLM Aligning Process in Figure 1) towards the decision-making preferences of one or more of the communities, or even individuals represented in the study. It is not known which of the large armamentariumWang et al. (2024) of alignment techniques would be most effective. To that end, a standardized set of metrics of compliance with alignment (e.g. the Alignment Compliance IndexKohane (2024) as well as reliability and robustness measuresGoto et al. (2024)) should be adopted to guide the implementation decisions for clinical applications of AI.

Third, and in parallel, the HVP will entail extensive societal discussions about which preferences we should be able to steer our AI clinical assistants (Normative Decision Guidance in Figure 1). It might be that we place primacy on individual preferences. However, often considerations other than the preferences of clinicians or patients may enter into decisions. For example, where there are limited resources, such as organs available for transplant, societal considerations such as those articulated by Persad et al.Persad et al. (2009) may be considered as normative.

The governance of the HVP will be of necessity multidisciplinary. Although, as diagrammed in Figure 1, AI researchers will be attending to the technical details of alignment measurement and aligning procedures, the values being assesses in the HVP are of broad human relevance and therefore all the participants in the Human Decision Guidance process must be represented in the governance structure.

We are already in the era of use of LLMs in clinical care, much of it driven by the interest and needs of individual clinicians and patients. As the use of these powerful cybernetic intellectual extenders rapidly encompass more of human decision-making and in healthcare in particular, the questions addressed by the HVP become central to human welfare, freedom and happiness. Fortunately, much preliminary work is already underway across multiple domainsHu et al. (2024); Eigner and Händler (2024) and leaders in medical AI increasingly recognize the urgency of embracing the ethical dimension in AI implementationsChar et al. (2018). The HVP is a response to that sense of urgency by recognizing that only a public, transparent and highly pragmatic representation of the wide variety of human values driving (clinical) decision-making and systematic studies of the alignment compliance of AI agents making their way into human decision making can ensure our individual and public good.

References

Maryam Bahrampour, Joshua Byrnes, Richard Norman, Paul A Scuffham, and Martin Downes. Discrete choice experiments to generate utility values for multi-attribute utility instruments: a systematic review of methods. *Eur J Health Econ*, 21(7): 983–992, 2020.

Danton S Char, Nigam H Shah, and David Magnus. Implementing machine learning in health care - addressing ethical challenges. *N Engl J Med*, 378(11): 981–983, 2018.

Ward Edwards. How to use multiattribute utility measurement for social decisionmaking. *IEEE Trans Syst Man Cybern*, 7(5):326–340, 1977.

Emanuel Eigner and Thomas Händler. Determinants of llm-assisted decision-making. *arXiv preprint*, 2024. Available from: http://arxiv.org/abs/2402.17385.

Samuel G Finlayson, Adarsh Subbaswamy, Karandeep Singh, et al. The clinician and dataset shift in artificial intelligence. *N Engl J Med*, 385(3):283–286, 2021.

Tomohiro Goto, Kazuki Ono, and Akira Morita. A comparative analysis of large language models to evaluate robustness and reliability in adversarial conditions. Technical report, Techrxiv, 2024. Available from: https://www.techrxiv.org/doi/pdf/10.36227/techrxiv.171173447.70655950.

Zhicheng Hu, Yiyang Ren, Jianlin Li, and Yue Yin. Viva: A benchmark for vision-grounded decision-making with human values. *ArXiv preprint*, 2024. Available from: http://dx.doi.org/10.48550/arXiv.2407.03000.

Isaac Kohane. Systematic characterization of the effectiveness of alignment in large language models for categorical decisions. *arXiv preprint*, 2024. Available from: http://arxiv.org/abs/2409.18995.

Michelle M Mello and Sherri Rose. Denial-artificial intelligence tools and health insurance coverage decisions. *JAMA Health Forum*, 5(3):e240622, 2024.

Govind Persad, Alan Wertheimer, and Ezekiel J Emanuel. Principles for allocation of scarce medical interventions. *Lancet*, 373(9661):423–431, 2009.

Rachel B Ramoni, John J Mulvihill, David R Adams, et al. The undiagnosed diseases network: Accelerating discovery about health and disease. *Am J Hum Genet*, 100(2):185–192, 2017.

Utility Research Conference on Subjective Probability and Decision Making. *Utility, probability, and human decision making: Selected proceedings of an interdisciplinary research conference, Rome, 3-6 September, 1973*. Kluwer Academic, Dordrecht, Netherlands, 1975.

Kimberly Splinter, David R Adams, Carlos A Bacino, et al. Effect of genetic diagnosis on patients with previously undiagnosed disease. *N Engl J Med*, 379 (22):2131–2139, 2018.

Anne M Stiggelbout, Trudy Van der Weijden, Maarten PT De Wit, et al. Shared decision making: really putting patients at the centre of healthcare. *BMJ*, 344(jan27 1):e256, 2012.

Peter Szolovits and Stephen G Pauker. Categorical and probabilistic reasoning in medical diagnosis. *Artif Intell*, 11(1–2):115–144, 1978.

Detlof Von Winterfeldt and Gregory W Fischer. Multi-attribute utility theory: Models and assessment procedures. In *Utility, Probability, and Human Decision Making*, pages 47–85. Springer, Dordrecht, 1975.

Zheng Wang, Bin Bi, Sai Krishna Pentyala, et al. A comprehensive survey of llm alignment techniques: Rlhf, rlaif, ppo, dpo and more. *arXiv preprint*, 2024. Available from: http://arxiv.org/abs/2407.16216.

Kun-Hsing Yu, Emma Healey, Tze-Yun Leong, Isaac S Kohane, and Arjun K Manrai. Medical artificial intelligence and human values. *N Engl J Med*, 390(20):1895–1904, 2024.

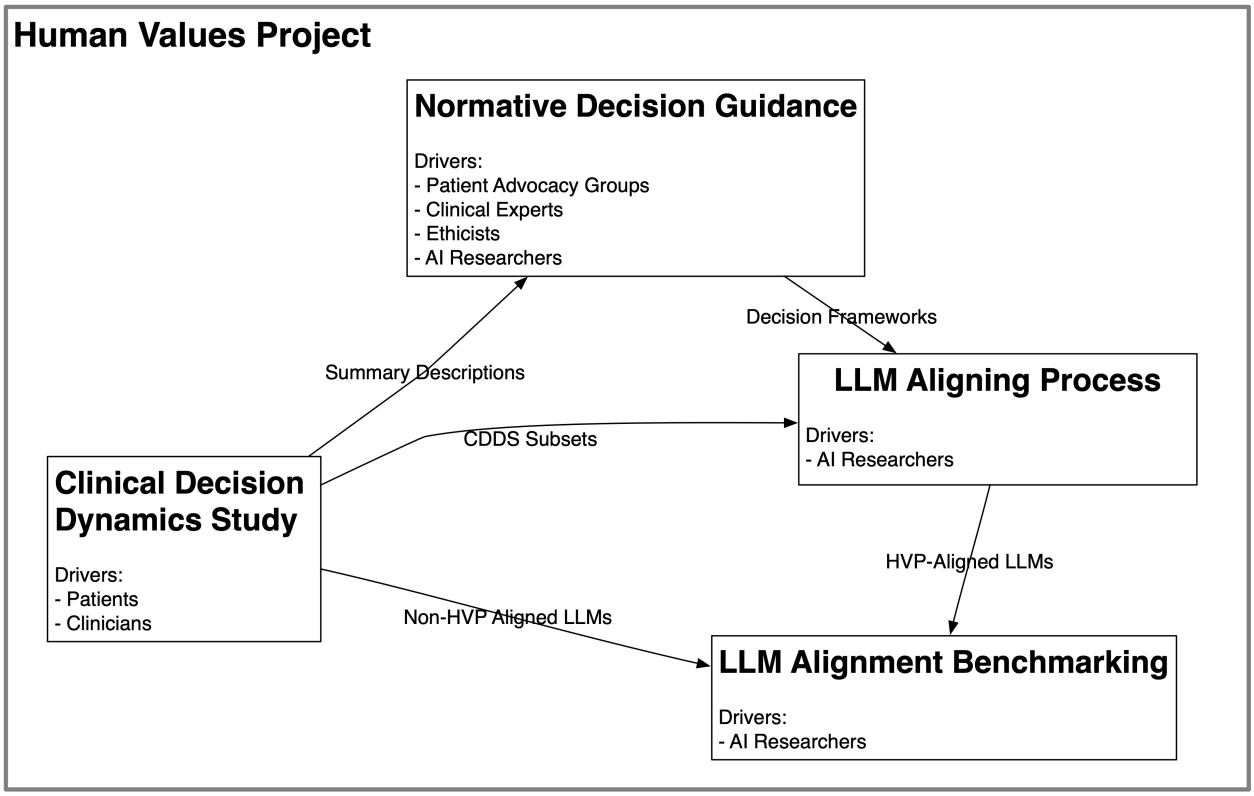

Figure 1: The HVP starts with the development of a Clinical Decision Dynamics Study (CDDS) driven by thousands of decisions by thousands of patients and clinicians. The CDDS is used as one of several perspectives informing the groups that are part of the Normative Decision Guidance process (patient, clinician, ethicists, AI researchers) developing decision frameworks. The CDDS human data are also used by the LLM Alignment Benchmarking processes (driven by AI researchers) to measure the alignment of those LLM's available without prior HVP guidance (Non-HVP Aligned LLMs). CDDS is also used for some of the alignment procedures (by the LLM Aligning Process, also driven by AI researchers). The decision frameworks produced by the Normative Decision Guidance processes help parameterize the LLM Aligning Process, sometimes also using the CDDS data subsets. The HVP-Aligned LLMs generated by the LLM Aligning Process are evaluated by the LLM Alignment Benchmarking process, where their alignment can also be compared to that of the Non-HVP Aligned LLM's.