Intelligent Health Systems – From Technology to Data and Knowledge
E. Andrikopoulou et al. (Eds.)

doi:10.3233/SHTI250639

Explainable Versus Interpretable AI in Healthcare: How to Achieve Understanding

Denis MOSER[a] and Murat SARIYAR[a,1]
[a]*Bern University of Applied Sciences, Switzerland*
ORCiD ID: Murat Sariyar https://orcid.org/0000-0003-3432-2860

Abstract. The increasing adoption of deep learning methods has intensified the demand for explanations regarding how AI systems generate their results. This necessity originated primarily in the domain of image processing and has expanded to encompass the complexities of large language models (LLMs), particularly in medical contexts. For example, when LLM-based chatbots provide medical advice, the challenge lies in articulating the rationale behind their recommendations, especially when specific features may not be identifiable. This paper explores the distinction between explanation, interpretation, and understanding within AI-driven decision support systems. By adopting Daniel Dennett's intentional stance, we propose a methodology for analyzing how AI explanations can facilitate deeper user engagement and comprehension. Furthermore, we examine the implications of this methodology for the development and regulation of medical chatbots.

Keywords. Explainable AI (XAI), Large Language Models (LLMs), chatbots, interpretation, understanding.

1. Introduction

With the rise of Deep Learning methods, the demand for explanations of how results are generated has intensified [1]. Initially, this need was particularly relevant in the field of image processing [2], but it has now expanded to encompass understanding the mechanisms behind the often-remarkable outcomes of LLMs [3]. For example, in tumor documentation using radiographic images, researchers seek to identify which human-interpretable features are responsible for tumor detection and characterization [4]. Similarly, in applications of LLMs – such as when a chatbot provides medical advice – the challenge of explanation becomes more intricate. Unlike image-based models, where explanations often point to visual features, LLMs rely on high-level abstractions derived from associating input with responses. Explanations in such cases often reference conceptual metaphors or patterns in such associations, leaving end-users with lingering questions about whether these insights truly foster understanding.

This raises a critical challenge: What does it mean to explain something, especially when the term "explanation" encompasses a variety of tasks across different contexts? Furthermore, does an explanation necessarily lead to understanding, and what factors must be considered to ensure that explanations are meaningful to specific users? To

[1] Corresponding Author: Murat Sariyar, Bern University of Applied Sciences, Quellgasse 21, CH2502 Biel/Bienne, Switzerland; E-mail: murat.sariyar@bfh.ch.

address these questions, this paper presents an approach for creating explanations that are tailored to specific practical scenarios. Unlike traditional methods, which often focus on generating feature-based explanations (e.g., techniques such as LIME [5] or SHAP [6]), this approach emphasizes bridging the interpretability gap—the distance between the technical explanation provided by an AI system and the user's ability to meaningfully understand and apply that explanation [7].

For example, while it is possible to explain in a deductive-nomological manner how a neural network is trained and how it generates a probability distribution over all possible classes, this does not provide insight into which specific features of the input were decisive for the final outcome, nor why the network processes information in the way it does. Genuine understanding requires not only an explanation but also the satisfaction of additional conditions that go beyond the explanatory mechanisms themselves. The proposed method is grounded in the following steps: (1) Defining practical use cases where interpretability is critical, such as LLM-based medical decision systems. (2) Evaluating existing explanation methods through the lens of user-centered factors like contextual relevance, cognitive accessibility, and trustworthiness. (3) Leveraging philosophical insights, such as Dennett's intentional stance, to frame explanations in ways that align with user-specific needs and decision-making processes.

2. Methods

The use case considered here focuses on the implementation of chatbots powered by LLMs for medical decision support. Such a system interacts with both patients and healthcare professionals, providing medical advice and aiding in diagnostics. The chatbot should ensure that its responses enhance the user's understanding of the reasoning behind the advice. For instance, when recommending a treatment protocol for diabetes management, an LLM might cite general medical guidelines. However, it should also link these recommendations to specific clinical trials, patient studies, or pathophysiological mechanisms that validate the suggested approach.

A clear distinction must be made between (i) explanation, (ii) interpretation, and (iii) understanding [8]. Explanation refers to the articulation of mechanisms or processes leading to a specific decision or outcome [9]. For example, in diagnostic support systems, an explanation might identify the symptoms or biomarkers that contributed to a particular diagnostic recommendation. Interpretation, on the other hand, involves tailoring the explanation to suit the audience's context or needs. A common approach, such as model distillation, simplifies complex predictive models (e.g., neural networks) into more interpretable forms, like decision trees, to make predictions comprehensible. Understanding denotes a cognitive achievement characterized by an individual's ability to synthesize and apply coherent knowledge effectively, particularly in clinical practice [10]. For instance, understanding the pathophysiology of heart failure requires knowledge of the mechanisms underlying fluid overload and reduced cardiac output.

Statements that contribute to a deeper understanding of a subject need not always be strictly factual. This principle underlies the power of metaphors in enhancing comprehension. A notable example is Daniel Dennett's concept of the intentional stance, which involves interpreting the behavior systems as if they possess beliefs, desires, and intentions [11]. In healthcare, metaphors play a pivotal role in bridging the gap between complex medical concepts and patients' understanding. By presenting information through metaphorical frameworks, practitioners can make intricate ideas more relatable

and accessible, facilitating a more meaningful grasp of conditions and treatments. While factual accuracy provides the foundation for medical education and communication, metaphors enrich understanding by offering intuitive, relatable contexts that clarify complexity [12]. However, the effective use of metaphors requires careful calibration. Practitioners and educators must balance metaphorical explanations with factual precision, ensuring that patients and students grasp the content's depth and implications while avoiding oversimplification or misconceptions [13].

3. Results

The primary advantage of Dennett's intentional stance, particularly in the context of LLMs, lies in its deliberate abstraction from the underlying mechanisms. This approach is both pragmatic and functional, as probing the internal workings of LLMs – such as attention-based next-token prediction – often yields descriptive explanations that fail to illuminate why these mechanisms produce advanced inferential capabilities. By focusing on the intentional stance, we embrace "useful fictions" that facilitate the evaluation of chatbot behavior. For instance, consider an LLM-based chatbot's response to a patient inquiry about managing diabetes (level of interpretation):

Chatbot: "*Based on your symptoms, I recommend a treatment protocol that includes dietary modifications, regular exercise, and monitoring your blood glucose levels. This aligns with established guidelines from the American Diabetes Association*."

The chatbot's response can be interpreted as reflecting beliefs, desires, and intentions. It "believes" the patient's symptoms suggest the need for specific interventions, based on learned patterns in diabetes management. It "desires" to improve the patient's health and "intends" to align its recommendations with evidence-based guidelines. This framing goes beyond simple explanation, offering an interpretation of the chatbot as a rational agent providing personalized, guideline-driven advice, which helps make the interaction more intuitive for the user. However, the response does not allow an interpretation of the model's internal reasoning process. Achieving a true interpretation would require a deeper exploration of why these interventions are proposed – specifically, how the model processes the patient's symptoms and synthesizes the information to generate a response. For instance, understanding entails elucidating how the model identifies the most relevant symptoms for a treatment recommendation and integrates this information with clinical guidelines:

Chatbot: "*Your reports of frequent fatigue and high blood sugar levels suggest that incorporating a high-fiber diet and increasing physical activity could help stabilize your blood glucose levels, as highlighted in recent clinical studies*".

For understanding, the chatbot must consider both the patient's level of prior knowledge as well as her emotional state. For example, if a patient exhibits limited understanding and expresses specific concerns, the chatbot might respond with:

Chatbot: "*I understand that managing diabetes can feel overwhelming. Let's break it down together. Increasing your fiber intake can help you feel fuller for longer, which may reduce cravings for sugary snacks and improve blood sugar control. Think of fiber as a speed bump for sugar – it slows the release of sugar into your bloodstream, giving your body more time to process it*."

This response elucidates the rationale behind the recommendations while addressing the patient's emotional concerns and need for metaphorical framing, thereby fostering a more empathetic and supportive interaction. However, for audiences such as physicians,

this approach may be less effective, as their priority lies in a detailed analysis of the underlying evidence.

In summary, Dennett's intentional stance offers a valuable perspective for enhancing interpretation and understanding in the context of LLM-powered medical chatbots. It underscores the importance of generating responses that are not only factually accurate but also emotionally sensitive and cognitively engaging, thus bridging the gap between information delivery and meaningful comprehension. Whether interpretation or understanding is required should be determined dynamically by the chatbot through the interaction itself. Achieving this capability may involve refining the chatbot via instruction tuning to enable it to intuitively adopt the intentional stance.

4. Discussion and Conclusions

In the ongoing discourse on the classification of medical products, particularly medical chatbots, this proposed layered approach offers a valuable basis for mapping chatbot functionalities to regulatory classes, guided by the European Medical Device Regulation, where intended use is decisive [14]. Class I medical devices, which involve minimal regulatory oversight, primarily focus on basic explanation and informational guidance. For instance, a chatbot operating within this class might simply outline symptoms associated with various conditions, offering general, low-risk information without engaging in clinical consultations or detailed justifications. As the chatbot's functionality advances to include the consideration of individual circumstances and the provision of interpretive guidance, it aligns with Class II devices, which are subject to moderate regulatory oversight. At this stage, the chatbot may provide personalized advice that, while helpful, introduces the potential for moderate risk – such as inappropriate recommendations [15]. Finally, when a chatbot's guidance involves high-risk scenarios, it would fall under Class III devices, which are subject to the most stringent regulatory scrutiny. A pertinent example is a chatbot advising a palliative diabetes patient in a manner that could lead to harmful self-management strategies, such as improperly adjusted insulin dosages. This mapping illustrates how the complexity and potential risks associated with interpretation and understanding in chatbot interactions correspond to increasing levels of regulatory classification.

To maximize the practical relevance of the proposed method for explanation, interpretation and understanding, its applications should be mentioned in clinical and regulatory contexts. For clinicians, this approach could guide the development of chatbots capable of delivering both accurate medical information and empathetic communication, improving patient interactions. For regulatory bodies, it offers insights into designing and evaluating AI systems that align with ethical standards and safety requirements. By conceptualizing chatbots and other AI systems as entities with beliefs, desires, and intentions, we can facilitate a more intuitive understanding of their functionality. This approach encourages users to engage with the technology more critically and thoughtfully, prompting them to consider the reasoning behind the AI's recommendations rather than merely accepting outputs at face value. However, while the intentional stance provides a compelling framework for facilitating user understanding, it may not be always sufficient in practical applications. Users often require concrete, factual information alongside metaphorical representations to make informed decisions about their health. The reliance on metaphors and narratives must be complemented by facts about the underlying algorithms and evidence that support the recommendations

provided by AI systems. Balancing relatable narratives with factual information is crucial to providing users with a comprehensive understanding of their options that extends beyond mere positive sentiment. Beyond the intentional stance, various metaphors have emerged in the discourse surrounding AI. One prominent metaphor is the concept of AI as a "black box", which underscores the opacity inherent in many machine learning models and highlights the challenges associated with comprehending their decision-making processes. Another salient metaphor is that of AI as an "augmented assistant", which suggests a collaborative relationship in which AI enhances human capabilities. This framing positions AI interactions as partnerships rather than as replacements for human roles. These metaphors serve distinct purposes and can be strategically employed to enhance user engagement, contingent upon the specific context in which AI is utilized. The aim is not solely to foster acceptance but also to evoke particular attitudes towards the technology. Looking ahead, future research should focus on exploring and refining the effectiveness of various metaphorical frameworks in different AI applications. Studies could investigate how users respond to various representations, such as the intentional stance versus more technical descriptions, in order to identify best practices for enhancing user comprehension and trust. Additionally, examining the long-term effects of these interactions on user behavior and health outcomes could yield valuable insights into the most effective ways to integrate AI into healthcare settings. As AI continues to evolve, understanding the interplay between metaphor, explanation, and user experience will be crucial for developing systems that truly empower users in their health management journeys.

References

[1] Longo L, Brcic M, Cabitza F, et al. Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. Inf Fusion 2024; 106: 102301.
[2] Borys K, Schmitt YA, Nauta M, et al. Explainable AI in medical imaging: An overview for clinical practitioners – Beyond saliency-based XAI approaches. Eur J Radiol 2023; 162: 110786.
[3] Zhao H, Chen H, Yang F, et al. Explainability for Large Language Models: A Survey. ACM Trans Intell Syst Technol 2024; 15: 20:1-20:38.
[4] Nazir S, Dickson DM, Akram MU. Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks. Comput Biol Med 2023; 156: 106668.
[5] Parisineni SRA, Pal M. Enhancing trust and interpretability of complex machine learning models using local interpretable model agnostic shap explanations. Int J Data Sci Anal 2024; 18: 457–466.
[6] Shirley ME, Kasujja NH, Marvin G. Shapley Additive Explanations (SHAP) for Cardiovascular Diseases Prediction. In: 2024 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS), pp. 1429–1437.
[7] Tzelves L, Kapriniotis K, Feretzakis G, et al. ChatGPT in Clinical Medicine, Urology and Academia: A Review. Arch Esp Urol 2024; 77: 708–717.
[8] Erasmus A, Brunet TDP, Fisher E. What is Interpretability? Philos Technol 2021; 34: 833–862.
[9] Duell J, Fan X, Burnett B, et al. A Comparison of Explanations Given by Explainable Artificial Intelligence Methods on Analysing Electronic Health Records. In: 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI). 2021, pp. 1–4.
[10] Elgin CZ. From Knowledge to Understanding. In: Hetherington SC (ed) Epistemology futures. Oxford University Press, 2006, pp. 199–215.
[11] Dennett DC. Taking the intentional stance seriously. Behav Brain Sci 1983; 6: 379–390.
[12] Nguyen J. Do fictions explain? Synthese 2021; 199: 3219–3244.
[13] Ross L. The Truth About Better Understanding? Erkenntnis 2023; 88: 747–770.
[14] Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20). Lancet Digit Health 2021; 3: e195–e203.
[15] Gilbert S, Harvey H, Melvin T, et al. Large language model AI chatbots require approval as medical devices. Nat Med 2023; 29: 2396–2398.