Can large language models help solve the cost problem for the right to explanation?

Lauritz Munch , Jens Christian Bjerring

Philosophy and History of Ideas, Aarhus Universitet, Aarhus, Denmark

Correspondence to
Dr Lauritz Munch;
lauritzmunch@gmail.com

Received 16 November 2023
Accepted 14 August 2024

ABSTRACT

By now a consensus has emerged that people, when subjected to high-stakes decisions through automated decision systems, have a moral right to have these decisions explained to them. However, furnishing such explanations can be costly. So the right to an explanation creates what we call the cost problem: providing subjects of automated decisions with appropriate explanations of the grounds of these decisions can be costly for the companies and organisations that use these automated decision systems. In this paper, we explore whether large language models could prove significant in overcoming the cost problem. We provide an initial case for believing that they can but only with serious ethical costs.

INTRODUCTION

Many claim that individuals affected by high-stakes decisions made by automated systems have a moral right to an explanation of these decisions.[1–6] For instance, when an AI (Artificial Intelligence) system is used to diagnose medical patients, these patients may be entitled to an explanation of why the system recommended a particular diagnosis. However, this claim faces challenges. While one problem concerns the well-known black-box problem,[7] we will here focus on a different and less discussed problem that we call the cost problem. Essentially, the problem is that it can be prohibitively costly to ensure patient understanding of AI decisions. If unresolved, the cost problem threatens to undermine the right to explanation.

Recent studies suggest that large language models (LLMs) may play a central role in alleviating the cost problem.[8] The key idea is to incorporate LLMs into various explainable AI (XAI) methods to create a dialogue partner that promotes understanding. We provide a case for believing that LLMs could reduce the need for human involvement in the process of explaining automated decisions. Yet, we conclude, that opting for this solution is associated with important ethical challenges.

How the right to explanation generates a cost problem

Why think there exists a moral right to explanation? One reason concerns autonomy. If you have no way of understanding why you were diagnosed in a certain way or recommended a particular medical treatment, it will be much harder to make informed decisions about your health. By fostering understanding, explanations may thereby support autonomy.[1] A second reason concerns the need to identify inaccuracies and unfairness in medical treatment. Understanding the reasons behind automated decisions allows individuals to detect and contest errors or biases, ensuring more fair and accurate outcomes ([2]; for criticism.[5 6]

Of course, not all medical decision-making processes warrant explanation. Examples may include specific clinical circumstances such as emergency situations where immediate action is necessary, or cases involving patients who lack the capacity to understand the explanation. While this might be true, however, it remains desirable to require explanations for medical decisions, whether made by AI systems or human doctors. Notably so when it comes to algorithmic decision-making, as algorithmic decision-making is often notoriously intransparent when compared with most human decision-making. Moreover, since medical algorithms typically are trained on vast datasets, they can contain biases and errors that are undetectable by humans.[9] While explanations of these algorithms may not reveal all these biases and errors, they at least increase the chance that we will identify and eliminate them.

Let us grant this initial case for a right to explanation. Which kind of explanation is then relevant? Ideally, we want explanations that elucidate which factors made a difference in the outcome of an automated decision and how they mattered. Examples of such explanations could be causal and difference-making explanations.[10] For instance, to foster an understanding of why a machine learning system predicts a high risk of pancreatic cancer for a patient, a relevant causal explanation might pinpoint the key segments of input data that made the biggest difference in the prediction.[11] Yet, even without causal explanations, various non-causal types of explanations may help us. In particular, there is hope that tools from the XAI toolkit may play many of the same roles as causal explanations when it comes to understanding the difference-making properties of automated decision systems.[11] Feature importance analysis, for example, may help identify which features in an input space have the most significant impact on an algorithm's predictions, whereas SHapley Additive exPlanations (SHAP) may provide an even more detailed view by quantifying how much impact each feature has on the algorithm's output.[12]

But explanations should not only tell us which factors made a difference in an automated decision. Ideally, they should also be personalised and tailored to individual patients. Whereas a technically trained person may be able to interpret a SHAP summary plot—a visualisation tool used to interpret the output of machine learning algorithms by showing the impact of individual features on the algorithm's predictions—most people will not. So explanations must be adapted to an individual's

To cite: Munch L, Bjerring JC. *J Med Ethics* Epub ahead of print: [please include Day Month Year]. doi:10.1136/jme-2023-109737

level of understanding of the concrete medical domain. Additionally, explanations should only provide patients with the most relevant information. If not, there is a risk of information overload that may restrict the patient's ability to make an informed decision.[13]

Providing subjects of automated decisions with explanations that meet these different requirements can obviously be a costly affair. Not only are there initial costs associated with developing systems that promote the explainability of automated decisions. For instance, developing, implementing and maintaining an XAI solution involves substantial costs related to advanced computing resources, specialised personnel and ongoing updates. Moreover, there are variable costs associated with delivering explanations to end-users that cater to their individual needs and abilities. When it comes to information from a SHAP summary plot, for example, there is no mechanical way to translate information from the summary plot into something that is useful for individual patients.[8] This creates the need for staffing human specialists with relevant domain expertise who can interpret the technical SHAP values and contextualise them in a manner that is comprehensible to each patient. Without such mediating specialists, XAI methods are likely to misfire and result in either information overload, misunderstanding or no understanding.[1]

Let us collect these different cost-increasing features of providing suitable explanations of automated decisions under the label of the cost problem. This problem not only threatens to diminish or even erode the net benefit of adopting automated decision systems in medical decision-making. It also threatens to undermine the case for thinking that there exists a right to explanation.[1] The main driver of the cost problem, we claim, is to be found in the variable cost of the human labour, which is needed to provide suitable explanations of automated decision systems to patients. After all, only humans have—at least so far—been able to effectively understand and convey the nuances involved in personalised and context-sensitive explanations.

How LLMs may solve the cost problem

Recent developments in LLMs and generative AI hold the promise of automating many tasks that were previously thought to be possible only for humans to accomplish. Focus on text production. Generating scripts, writing articles, crafting poems and engineering code used to be a job exclusively for humans. So did tailoring writing styles and answers to specific audiences and genres. And so did collaborating with human writers on creative and intellectual projects. But all these tasks can now be approached with great success by LLMs.[14–16]

Likewise, it seems possible to train LLMs to act as mediating specialists, translating various XAI explainability methods into explanations that promote patient understanding. Here is how it might work. Suppose a hospital deploys a complex black-box algorithm to differentiate between malignant and benign tumours, considering factors such as tumour size, shape, density and patient medical history. Due to the nature of the algorithm, patients may not understand why they receive a 'malignant' or 'benign' diagnosis. To remedy this situation, the hospital decides to implement a SHAP-based model on top of the opaque diagnostic algorithm, with the resulting SHAP values elucidating how much features like tumour size and density influence the diagnosis compared with patient age and genetic markers. As mentioned, summaries of such SHAP values might not be meaningful to non-specialists. Recognising this, the hospital decides to fine-tune an LLM to interpret the summary plot for patients. Initially, they gather a diverse set of diagnostic cases containing relevant input features such as tumour characteristics and patient history. They also collect the SHAP values associated with each case and the diagnostic algorithm's predictions (malignant/benign). A team of medical experts then provides a range of explanations for the algorithmic diagnosis using the input features and SHAP values, tailored to hypothetical patient inquiries. Ideally, these explanations will span widely, catering to various informational needs and comprehension levels.

Once all this data is compiled, the hospital uses supervised learning to train the LLM by supplying it with sets of input data, including the relevant medical features, SHAP values, opaque algorithmic prediction and a specific query, along with the corresponding expert-generated explanation as the target output. Through this training, the LLM learns to discern the connections between the input data and the expert explanations. Assuming that the training data encompass a wide enough array of inquiries and explanations, the LLM should learn to adapt its explanations to diverse patients. Depending on the patient's inquiry, the LLM might emphasise which features most significantly impact the algorithmic diagnosis, whether they contribute positively or negatively to the diagnosis, and how these features relate to the individual's case. Ultimately, a patient might receive an LLM-generated explanation along the lines:

'While the diagnosis indicates a benign tumour, it is important to understand which factors led to this conclusion. The size and density of the tumour were key considerations, while your family's medical history played a minor role in the assessment. Your overall health and recent medical tests also contributed positively to the diagnosis'.

Since LLMs in this way can learn to personalise explanations based on specific users and queries, they have the potential to play the role of the mediating specialists that we characterised above.

Slack *et al* have documented a proof of concept of these ideas that they call TalkToModel[8]: 'a system that enables open-ended natural language dialogues for understanding (machine learning) models for any tabular dataset and classifier'.[7] While we shall not go into the details of this system here, what matters for our purposes is that Slack *et al* make use of an LLM to mediate between the questions, which users may have about an automated decision affecting them, and the answers that they receive from the interpretability model associated with the relevant opaque algorithm. Based on this natural language interface, Slack *et al* claim that the TalkToModel has a number of desirable features. Not only can the model facilitate discussions about why specific algorithmic predictions were made, and about how these predictions may change if the input data were to change. It can also answer questions about how the opaque model works in general as well as supporting end-users by asking follow-up questions.[8]

As such, it is not merely a theoretical possibility that LLMs may help mitigate the cost problem. If LLMs can learn to accomplish many of the communicative roles that could previously only be handled by human experts—by experts who are able to understand and effectively communicate the findings of explainability models like SHAP to end-user—LLMs can also help reduce the costs involved in providing explanations of automated decision systems. Moreover, since LLMs will be able to handle large volumes of explanations simultaneously, they also have the potential to scale up in ways that human labour cannot, thereby further increasing cost savings. Finally, since LLMs can continuously learn and improve with ongoing use, they can improve over time to deliver increasingly adequate explanations to users, thereby decreasing the time and the costs associated with generating explanations to users.

Taken together, then, it seems that LLMs can play a central role in overcoming the cost problem: they can deliver explanations that foster suitable understanding without incurring the high costs associated with human labour.

Ethical challenges from using LLMs to solve the cost problem

Yet, if we attempt to use LLMs to overcome the cost problem for the right to explanation, there are certain distinct ethical concerns that we need to be aware of.

Risk of discrimination

We have argued that LLMs can help translate technical explanations into simpler explanations that may foster understanding in patients and cater to their diverse needs. This process, however, introduces risks of discrimination and unequal treatment. Notably, because patients have to express their explanatory needs to the LLM via natural language prompts. Even if we set aside concerns about technical skills in prompting effectively, it is well documented that capacities and expertise in articulating one's needs through natural language are unequally distributed.[17] On average, people with a higher socioeconomic status are better off in this regard. This creates the risk that using LLMs to solve the cost problem may inadvertently sustain, if not exacerbate, already existing inequalities in healthcare due to indirect socioeconomic discrimination.

While this risk is significant, alternatives to LLMs are beset with quite similar problems. For instance, direct and indirect socioeconomic discrimination also happens in interactions with human doctors, as they too can be influenced by class-based biases.[18] Of course, doctors may be in a better position than LLMs to correct some of their biases when interacting with patients. Unlike LLMs, doctors are not confined to interacting with patients through ordinary language; instead, they can often gain much information about a patient's explanatory needs by observing their non-verbal behaviour.[19–22] For example, a doctor might notice a patient's gaze and discern that they are confused or anxious about a particular aspect of their diagnosis. Such non-verbal cues can prompt the doctor to provide additional clarification or reassurance, tailored to the patient's immediate needs. In this sense, the ability to interpret non-verbal cues gives doctors a richer evidential basis for addressing the explanatory needs of patients. While so-called multimodal LLMs may eventually remedy this situation—as they have the potential to interpret non-verbal cues for users—at least for now a purely text-based LLM solution to the cost problem significantly reduces the number of ways in which explanatory needs can be articulated and met.

Risk of fabrication

Another risk stems from the fact that LLMs have a tendency to fabricate or 'hallucinate' information. For example, when asked about factual information, they may tell us about books that were never published, and about events that never happened.[23]

This is particularly worrisome in the medical domain. If LLMs are prone to offering false, fabricated or misleading explanations, they may mislead patients into thinking they genuinely understand what made an algorithm yield its prediction, when in fact they do not. In the worst case, false or misleading explanations may motivate patients to undertake courses of action that are either unwarranted or directly counterproductive for their overall health situation. Suppose, for instance, that an LLM explains to a patient that the primary cause of their disease is genetic disposition, when in fact it is primarily caused by the patient's lifestyle. Without accurate information, the patient might wrongly conclude that no lifestyle changes are needed, thereby missing the chance to take beneficial action.

Fabricating information is indeed a challenge when using LLMs for generating explanations based on XAI methods. Of course, it may be noted that human doctors are not perfect either. Occasionally, they also end up giving patients false explanations. However, this point may underestimate the extent to which LLMs are different from humans when it comes to the types of mistakes they make. First off, well-meaning human doctors simply do not fabricate medical data to arrive at a diagnostic recommendation. They may misunderstand the available data in different ways, but they simply do not produce it on a whim. Second, LLMs tend to err in ways that are both different from and harder to anticipate than the errors made by human doctors.[24] That is, while human experts typically make mistakes that follow certain recognisable patterns due to common human limitations or biases, the mistakes made by machine learning algorithms can be more random, more extreme and less foreseeable.

There is no easy fix to these problems. However, there are approaches that may help mitigate the issue in the highly fine-tuned and specific contexts where we imagine that LLM models are employed. First, we should attempt to ensure that training and fine-tuning data sets are of high quality: they need to be diverse and representative of various question-answer scenarios. This will help the LLM learn a wide range of accurate responses and thus reduce the likelihood of generating incorrect information. Second, we should attempt to incorporate various fact-checking mechanisms into the training process to catch and correct factual errors. We may, for instance, allow the LLMs to access peer-reviewed articles such as those that can be found on PubMed to cross-check information against these databases to verify the accuracy of their explanations.[23] Finally, we may try to reduce information fabrication by enhancing the fine-tuned LLMs themselves. For instance, we may attempt to integrate various rule-based systems with the LLMs to help them operate more strictly within well-defined parameters.

It also bears mentioning that human fact-checking accounts for a central part of the training phase for the LLMs. Medical experts initially curate and verify a diverse set of diagnostic cases, ensuring that the explanations provided by the LLMs during training are accurate and contextually appropriate. This work helps the LLMs learn how to generate reliable explanations based on the data provided by experts. In fact, it is compatible with what we have said so far that human fact-checking is involved also during the LLMs' inference phases. Of course, expecting humans to check every single LLM explanation would reintroduce many of the costs that the use of LLMs is meant to save. But human fact-checking can be employed more strategically.[25] For example, explanations that have a significant impact on patient care, involve high-stakes decisions or are flagged by automated systems for potential issues can be prioritised for human review. Additionally, regular audits of randomly selected explanations can help ensure overall quality and identify any systemic issues in the LLMs' outputs. In this way, by reserving human oversight for where it is most needed, we can expect to reduce the negative impact of information fabrication.

Risks to privacy

A third risk associated with using LLMs to overcome the cost problem stems from their tendency to 'leak' sensitive or confidential information present in their training data.[26] LLMs can also be susceptible to prompt injection attacks, where malicious users craft inputs designed to manipulate the models into generating unauthorised responses from which misleading information may then be extracted.[27]

It is worth highlighting explicitly that this problem is more serious for LLMs than for other methods of meeting a patient's need for explanations. For example, while a human practitioner could, and sometimes does, leak confidential information, this risk is much less pronounced. Not only can access to confidential information be restricted to specific human agents, but human practitioners are also bound by strict ethical guidelines and professional standards that emphasise the importance of patient confidentiality. Absent such ethical constraints and training, LLMs may, by contrast, inadvertently process and reveal confidential information, especially if not properly supervised or if their data handling protocols are not meticulously designed.

Again, these are serious challenges to the idea of using LLMs to overcome the cost problem. As above, however, there are various ways in which they could be addressed. We may, for instance, implement differential privacy techniques during training to make it difficult to reliably infer any individual data points from the LLM's output.[28] We may of course also restrict who has access to the LLMs and include various monitoring devices for their usage.

CONCLUSION

We have argued that LLMs may play a significant role in overcoming the cost problem for the right to explanation. With proper training, LLMs can deliver personalised explanations of opaque automated decisions to individual patients, making them both efficient and cost-effective due to their scalability. However, we should be careful not to paint too glamorous a picture of the potential of LLMs in offering explanations. As we have seen, they also bring significant ethical concerns that must be addressed, including LLM accuracy, discrimination and privacy. Moving forward, we need empirical data on how best to test and implement LLMs in automated decision-making in medicine. This involves designing rigorous validation protocols and pilot studies to assess the real-world effectiveness and safety of LLMs. Understanding user interactions and feedback will be crucial in refining these models to ensure they meet the needs and expectations of patients.

Contributors Each author contributed to all processes. LM is the guarantor for the overall content.

Funding This study was funded by Carlsbergfondet (CF20-0257).

Competing interests None declared.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data sharing not applicable as no datasets generated and/or analysed for this study

ORCID iDs
Lauritz Munch http://orcid.org/0000-0002-3510-5422
Jens Christian Bjerring http://orcid.org/0000-0001-8755-6746

REFERENCES

1 Vredenburgh K. The Right to Explanation. *J Pol Phil* 2022;30:209–29.
2 Lazar S. Legitimacy, authority, and democratic duties of explanation. In: Sobel D, Wall S, eds. *Oxford Studies in Political Philosophy*. . 2024: 10. 28–56.
3 Purves D, Davis J. Public Trust, Institutional Legitimacy, and the Use of Algorithms in Criminal Justice. *Pub Aff Q* 2022;36:136–62.
4 Grant DG, Behrends J, Basl J. What We Owe to Decision-Subjects: Beyond Transparency and Explanation in Automated Decision-Making. Philosophical Studies, 2023.
5 Taylor E. Explanation and the Right to Explanation. *J Am Philos Assoc* 2023;1–16.
6 Da Silva M. Explainability, Public Reason, and Medical Artificial Intelligence. *Ethic Theory Moral Prac* 2023;26:743–62.
7 Zednik C. Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. *Philos Technol* 2021;34:265–88.
8 Slack D, Krishna S, Lakkaraju H, *et al*. Explaining machine learning models with interactive natural language conversations using TalkToModel. *Nat Mach Intell* 2023;5:873–83.
9 Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, *et al*. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fus* 2020;58:82–115.
10 Baron S. Explainable AI and Causal Understanding: Counterfactual Approaches Considered. *Minds & Machines* 2023;33:347–77.
11 Molnar C. Interpretable Machine Learning. Lulu.com, 2020.
12 Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst* 2017;30.
13 Leichtmann B, Humer C, Hinterreiter A, *et al*. Effects of explainable artificial intelligence on trust and human behavior in a high-risk decision task. *Open Science Framework* [Preprint] 2022.
14 McKinsey and Company. The Economic Potential of Generative AI, 2023. Available: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier
15 Hernandez-Olivan C, Hernandez-Olivan J, Beltran JR. A survey on artificial intelligence for music generation: agents, domains and perspectives. *arXiv* [Preprint] 2022.
16 Olga A, Saini A, Zapata G, *et al*. Generative ai: implications and applications for education. *arXiv* [Preprint] 2023.
17 Der Nederlanden SJ, Schaeffer JC, Van Bakel HHJA, *et al*. 2023 Socio-economic status and other potential risk factors for language development in the first year of life. *J Child Lang*2023:1–21.
18 Rickett B, Easterbrook M, Sheehy-Skeffington J, *et al*. The British Psychological Society; Psychology of Social Class-Based Inequalities: Policy Implications for a Revised, 2022. Available: https://explore.bps.org.uk/binary/bpsworks/b5c9f3afe2f3b45b/c831f5664ba3cea5cfa8e9b372e809c81bd380dc0a801d18dd383b32b57f5abf/bpsrep_rep167.pdf
19 Lenharo M. Google AI has better bedside manner than human doctors — and makes better diagnoses. *Nature New Biol* 2024;625:643–4.
20 Tu T, Palepu A, Schaekermann M, *et al*. Towards conversational diagnostic AI. *arXiv* 2024.
21 Laranjo L, Dunn AG, Tong HL, *et al*. Conversational agents in healthcare: a systematic review. *J Am Med Inform Assoc* 2018;25:1248–58.
22 Abdulghafor R, Abdelmohsen A, Turaev S, *et al*. An Analysis of Body Language of Patients Using Artificial Intelligence. *Healthcare (Basel)* 2022;10:2504.
23 Béchard P, Ayala OM. Reducing hallucination in structured outputs via retrieval-augmented generation. *arXiv* [Preprint] 2024.
24 Alvarado R. Should we replace radiologists with deep learning? Pigeons, error and trust in medical AI. *Bioethics* 2022;36:121–33.
23 Mökander J, Schuett J, Kirk HR, *et al*. Auditing large language models: a three-layered approach. *AI Ethics* 2023;1–31.
26 Lukas N, Salem A, Sim R, *et al*. Analyzing leakage of personally identifiable information in language models. 2023 IEEE Symposium on Security and Privacy (SP; San Francisco, CA, USA, 2023
26 Freiesleben T. The Intriguing Relation Between Counterfactual Explanations and Adversarial Examples. *Minds & Machines* 2022;32:77–109.
28 Danger R. Differential Privacy: What is all the noise about? *arXiv* 2022.