ETHICS OF APPLYING LLM-MODELS IN MEDICINE AND SCIENCE

Gabidullina LF[1]✉, Kotlovsky MY[1,2]

[1] Yaroslavl State Medical University, Yaroslavl, Russia

[2] Department of Strategic Analysis in Healthcare N. A. Semashko National Research Institute of Public Health, Moscow, Russia

Rapid integration of large language models (LLM) into healthcare gives rise to acute ethical dilemmas and practical risks. The principal issue is associated with trust of medical professionals, patients and developers in the models, as well as with the potential violation of medical ethics. In the article, key challenges are analyzed including critical importance of trust (depending on LLM data quality), disturbance of informed consent and autonomy of a patient due to the lack of transparency and excessive trust in AI algorithms. Particular attention is given to the risks of confidential medical data protection, which is confirmed by non-authorized transfer of data while using generally accessible LLM. The need to develop transparent, safe and ethically regulated solutions for LLM is medicine is prioritized.

Key words: ethical dilemmas, large language models (LLM)

Author contribution: Gabidullina LF — literature analysis, research planning, writing, editing; Kotlovsky MYu — data collection, analysis, and interpretation.

✉ **Correspondence should be addressed**: Landush F Gabidullina,
Revolyutsionnaya str., 5, Yaroslavl, 150000, Russia; landush10@yandex.ru

Received: 23.07.2025 **Accepted:** 05.09.2025 **Published online:** 22.09.2025

DOI: 10.24075/medet.2025.016

ЭТИКА ПРИМЕНЕНИЯ LLM-МОДЕЛЕЙ В МЕДИЦИНЕ И НАУКЕ

Л. Ф. Габидуллина[1]✉, М. Ю. Котловский[1,2]

[1] Ярославский государственный медицинский университет, Ярославль, Россия

[2] Национальный научно-исследовательский институт общественного здоровья имени Н. А. Семашко, Москва, Россия

Быстрое внедрение больших языковых моделей (LLM) в здравоохранение порождает острые этические дилеммы и практические риски. Центральная проблема связана с доверием медицинских специалистов, пациентов и разработчиков к этим системам, а также с потенциальным нарушением основополагающих принципов медицинской этики. Данное изложение анализирует ключевые вызовы, включая критическую важность доверия (зависящую от качества данных LLM), нарушение информированного согласия и автономии пациента из-за отсутствия прозрачности и чрезмерной опоры на ИИ. Особое внимание уделяется рискам защиты конфиденциальных медицинских данных, что подтверждается инцидентами несанкционированной передачи информации при использовании общедоступных LLM. Необходимость разработки прозрачных, безопасных и этически регулируемых решений для LLM в медицине становится первостепенной задачей.

Ключевые слова: этические дилеммы, большие языковые модели (LLM)

Вклад авторов: Л. Ф. Габидуллина — анализ литературы, планирование исследования, написание текста, редактирование; М. Ю. Котловский— сбор, анализ, интерпретация данных.

✉ **Для корреспонденции:** Ландыш Фаритовна Габидуллина
ул. Революционная, д. 5, г. Ярославль, 150000, Россия; landush10@yandex.ru

Статья поступила: 23.07.2025. **Статья принята к печати:** 05.09.2025 **Опубликована онлайн:** 22.09.2025

DOI: 10.24075/medet.2025.016

In our work, we frequently resort to digital technologies and artificial intelligence. But can we be sure that what we do in the digital world is always ethical and safe? We are convinced that integration of advanced technologies, and large language models (LLM) in particular, into healthcare and science requires not only technical knowledge, but also deep ethical understanding. This relevant topic is highlighted in our article.

Large language models (LLM) are a variety of artificial intelligence (AI) based on the transformer architecture and trained on vast quantities of text data to perform a wide range of tasks related to natural language processing (NLP), such as text generation, translation, summary, question-answering systems, coding, and others.

Language models can learn about themselves by introspection (for instance, they can predict the next word). They can be pre-trained on trillions of tokens from various sources (the Internet, books, scientific articles, etc.). The models have billions of parameters (for example, 175 billion parameters for GPT-3 and over 500 billion parameters for GPT-4). They can be guided to perform tasks through zero-shot learning, few-shot learning, and fine-tuning. The models produce a human-like text, but fail to understand it as a human does [1].

LLM models are trained to do the following:

- collection of data;
- tokenization (Byte Pair Encoding (BPE) or SentencePiece);
- pretraining;
- fine-tuning.

While dealing with LLM, the following stages are used:

- prompting;
- request encoding through torenization. Tokens are transformed into embedding vectors;
- inference. Vectors go through transformer layers. In a transformer layer, the self-attention mechanism allows each token to process the context of all other tokens. The model predicts the next token in a sequence by choosing the most probable one;
- response formation;

– post-processing and filtering (in production). Additional modules (retrievers, knowledge bases, etc.) can be used sometimes.

Speech recognition converts spoken words into text from audio files. Speech synthesis is still a difficult task, especially when it should sound natural and emotional.

Generation of images from text descriptions is aimed at the creation of a text-based image. The task is still rather complex. However, diffuse models or architectures used to generate images in the field of computer vision have been successfully implemented during the last years. They have gained particular popularity since 2020. In 2021, MidJourney neural network that uses diffusion models has been created, whereas in June 2022, Sber presented the Kandinsky neural network.

Both neural networks can generate text-based images of a very good quality.

Question-Answering Systems are machine learning models that can find answers to text-based questions. Thus, LLM have remarkable capabilities of text processing and generation, data analysis, and creating recommendations.

PATIENTS AND METHODS

There are more than forty different LLM models. Their funding is raising every year. In 2022, they surpassed average human capabilities in associative thinking. However, the potential is accompanied by new ethical dilemmas, which are frequently complex. In healthcare, where human health and life are at stake, these issues become particularly acute. The central aspect that requires close attention is trust of medical professionals, patients and developers in LLM-based tools, system that uses these technologies and reliability of these systems under critically important conditions respectively [2].

Trained on enormous amounts of data, often closed or insufficiently verified, lacking transparency and full explainability of algorithmic solutions, they cannot always explain their conclusion and be logical. In clinical practice, such non-transparency is a source of potential risk if the assistant's recommendations are accepted without being critically considered by the doctor. Can you trust a tool whose decisions are inexplicable? Ethical practice demands minimum transparency for any tool that affects human health.

Trust is closely interrelated with quality and representativeness of data used to train LLM. Medical data can be incomplete, distorted, biased or erroneous due to specifics of medical history taking, subjective interpretation of symptoms, regional differences in medical approaches or even social and economic status of patients. As LLM can see regular patterns in incorrect or biased data, recommendations that do not correspond to the best clinical practices or principles of medical ethics can be generated [3].

For instance, a model can be too general offering a standard solution, ignoring individual traits or concomitant diseases of a patient, which can be harmful.

The way how patients perceive LLM-assistants is equally important. If patients are not aware of using artificial intelligence (AI) models in their diagnosis or treatment recommendations, it proves that the principle of informed consent and autonomy has been violated. The ethical interaction imperative should be transparent: patients must be aware of artificial intelligence involvement in their treatment and have a right to reject it.

Apart from clinical practice, large language models are actively introduced into medical research. They are used to analyze an enormous set of scientific publications, generate hypotheses, support experiment design, summarize results and perform primary analysis of data during preclinical and clinical research. Wider application of LLM-assistants is associated with new ethical risks such as:

– authenticity and "hallucinations" (completely fictitious, but well-formed statements; incorrect recommendations (for example, medications with contraindications); false sources or references (non-existent scientific publications);
– biased research;
– issues of authorship and intellectual property;
– reproducibility and verification;
– data confidentiality.

Thus, ethical regulation requires an enhanced complex approach and interdisciplinary cooperation, namely:

– we need LLM capable of explaining the logics of their decisions;
– bias and errors in data used to train the models should be minimized;
– patients and research participants should be informed of using AI and have a choice;
– clearly define responsibilities of a developer, doctor who used the system or the system itself in case of an error;
– making recommendations on using LLM in science, including the issues of authorship, accuracy and reproducibility;
– medical professionals and scientists should know how to use LLM properly and critically, understand limitations and ethical aspects.

If the extensive training samples are based on historically developed prejudices, inaccuracies or systemic irregularities typical of real clinical practice, the models can adopt, replicate and aggravate them.

Let's take an ethnic prejudice as an example. If training data historically underestimated symptoms or frequently ignored the needs of patients from certain ethical groups, LLM can adopt and reproduce the undesirable patterns.

The same is about the gender-based shift. The male model has been the focus of medical research for a long time while female-specific needs were ignored.

Patients with health limitations, mental disturbances or representatives of vulnerable groups are also at risk. As a LLM can reproduce stereotypes in a conscious way, it can be a source of involuntary discrimination in clinical practice [4].

The most dangerous part concerns a concealed nature of these shifts. An LLM is not capable to explain the logics of its solutions due to non-transparency and lack of total explainability of algorithmic solutions mentioned before. The recommendations can be convincing and accurate producing a false feeling of objectivity of a doctor. Moreover, an algorithmic prejudice is rarely seen in single cases; it becomes obvious only when large sets of data are analyzed in aggregate. However, an erroneous decision can have irreversible effects for the patient.

Many legislative acts urge medical institutions to follow the highest standards of storing, processing, accessing and transferring data of patients. However, as soon as an independent technological system such as a LLM appears, the risks of potential disturbance of safety and confidentiality increase multiple times [5].

Breach of confidentiality can involve as follows:

– mental and social harm to a patient;
– loss of trust in healthcare;
– legal liability.

LLM can collect, process, reproduce or involuntary disclose the data directly or indirectly, especially when trained and used in medicine.

Many language models, especially the ones offered by commercial developers, can have inbuilt mechanisms of interaction logging. The fact is of particular concern. Even if the information is depersonalized, complex data correlations are associated with a high risk of reidentification of a patient. It is absolutely unacceptable for basic ethical principles of medical practice.

Can a medical professional trade absolute patient confidentiality for a more accurate and machine-based analysis of symptoms?

Patients should not only sign templates, they need to obtain clear, illegible and comprehensive information regarding what type of data can be used by LLM, where and how it is done, which potential risks are associated herewith, and who will get access to it. Ethical management of data goes far beyond formal legal protection.

Reliable technical and organizational protocols that ensure absolute data confidentiality and safety should not be a desirable condition only but also an essential requirement for their ethical implementation. Otherwise, even the most exact and potentially 'useful' information can become a source of deep violation of basic rights of a patient and undermine the trust we are trying to build.

Further to our discussion regarding trust, which is closely related to data transparency and protection, it is necessary to mention how an LLM influences a patient's autonomy and essence of medical ethics.

Information asymmetry and patient alienation: if a doctor will trust AI and fail to provide the patient with clear information, the patient will be marginalized while making decisions and deprived of a true informed decision.

Erosion of medical subjectivity and reliability: traditionally, a doctor does not only embody intellectual aspects, but is also guided by empathy, compassion and deep personal reliability. Despite powerful analytical capabilities, artificial intelligence is deprived of human qualities. Its recommendations are based on algorithms and statistics but not on ethical estimation or comprehension of a unique human situation. Extreme dependence of a doctor on LLM conclusions and their perception as an unquestioned objective authority can decrease a doctor's critical thinking and weaken the ethical position of the person who is ultimately responsible for making decisions.

Affecting a patient's trust in a doctor: it is directly associated with the covered topic of 'trust'. If the patient feels that key health-related decisions are taken by or strongly depend on the machine but not on a live doctor, the patient's belief in true care and individual approach will be undermined.

Limited choice and standardized decisions: there exists a risk that LLM, which are optimized to get the most 'optimal' or statistically substantiated recommendations, will substitute the 'individualized approach' with standard protocols. It is especially dangerous in case of automated triage or in a limited access to direct medical contact.

RESEARCH RESULTS

Misinformation provided by ChatGPT in response to medical questions

In a research (Ayers et al., 2023), responses of doctors and ChatGPT to real medical questions from patients have been compared. Though responses from AI sounded more polite, they contained potential dangerous or inaccurate data in 27% of cases.

Reaction: JAMA warning and calls not to use ChatGPT in telemedicine without verification [6].

Fake news and misinformation in telemedicine

Researchers tested the ability of GPT to generate misinformation. It was easy for the model to lie about the 'new virus', 'cancer-causing vaccine', etc.

Reaction: UN and WHO are calling for caution to be exercised in using AI in public healthcare without an ethical expertise [7].

Fake articles and sources in works of students and scientists

In 2023–2024, students around the globe submitted works containing AI-generated fake citations.

Result: universities introduced official regulations targeting the irrational use of LLM and took strict measures to combat plagiarism.

The University of Melbourne abolished a diploma due to imaginary sources being discovered in a master's dissertation [8].

В 2023–2024 годах во многих университетах мира студенты начали массово сдавать работы, содержащие вымышленные библиографические ссылки, сгенерированные LLM.

Patient data leakage via ChatGPT in Samsung (2023)

Employees of Samsung in South Korea used ChatGPT to process internal documentation including medical records and analysis of diagnostic code in biomedical software.

Issue: correspondence with LLM is preserved on OpenAI servers and can be used to educate the models if the respective settings were not disabled. There was a risk of leakage of sensitive data.

Result: Samsung prohibited to use public LLM data. The company started development an offline model of its own [9].

NHS of Great Britain: using GPT via third interfaces without verification

In some hospitals, doctors started using public web-versions of ChatGPT to generate effective discharge summaries, recipes and abstracts. They sometimes copied fragments from medical records to the chat interface.

Risks: data are sent to servers outside the jurisdiction of Great Britain (and GDPR) violating the laws on medical secrecy protection.

Results: NHS issued an urgent note not to use open LLM until protected decisions are implemented [10].

DISCUSSION OF RESULTS

What regulations should be followed while complying with the ethics of using technologies and AI in medicine?

1. The Man Above All principle. All decisions and developments should be done for the benefit of patients and to protect their dignity and rights, but not to enrich technological capabilities.
2. Transparency and openness of information regarding data collection, use and protection and regarding how and to

which extent AI participates in the process of diagnostics and treatment is the foundation for informed consent and trust.
3. It is necessary to determine clearly who is ultimately responsible for AI-based decisions. In medicine, it is always a doctor who should be responsible for the decisions.
4. Ethics of AI use requires to stop discrimination and bias that can be typical of algorithms and to ensure equal access to qualitative aid.
5. Technologies must provide patients with an informed choice instead of limiting it and encourage their active participation in health management.

Several key directions for further development are suggested:

1. Develop national and international ethical standards and guidelines that will allow using LLM in healthcare taking into account not only technological but also ethical aspects.
2. Integrate AI ethics in medical education and continuing professional development. Future and practicing doctors should be able to work not only with technologies but also with their ethical aspects.
3. Create multidisciplinary teams (doctors, ethicians, lawyers, engineers, representatives of patients) that constantly monitor, assess and adjust ethical principles to rapidly changing technology.
4. Prioritize research that examines AI long-term effect on doctor-patient relationships and psychoemotional condition of doctors and patients.
5. Actively implement joint decision making when information is submitted by AI, but it is a doctor-supported patient who makes a final decision.

Practical and ethical recommendations:

- prohibition to use open LLM to enter Personal Medical Data (PMD) without any special agreements;
- data anonymization prior to processing;
- local or protected LLM services within the clinic;
- development of ethical protocols of consent to AI-based data processing;
- mandatory auditability of AI use in medical IS;
- digitalization and logging of all LLM references while working with patients.

To sum it up, it is obvious that LLM implementation in healthcare is not a simple technological breakthrough but also a deep ethical challenge requiring a conscious and reliable approach. Thus, trust of patients in the system, doctor and, ultimately, the technology itself is a crucial point here. The trust cannot be achieved without strict adherence to the principles that have been discussed today.

CONCLUSIONS

Ethical issues with large language models (LLM) in medicine represent a complex and multilevel challenge. Issues of reliability, validity of information, liability for errors, confidentiality of data, bias and discrimination, transparency and explainability, as well as unequal access should be carefully analyzed and strictly regulated. It is an integrated approach, including technical, legal and social measures, that will reduce risks and expand the use of LLM potential in clinical practice.

On the way to the digital age of medicine we, therefore, should follow the light of innovations and ethical principles warranting those technologies serve for the benefit and health of humans, and not the other way around.

References

1. Kosarev Ye A. Uchebnik po mashinnomu obucheniyu. Available from: https://education.yandex.ru/handbook/ml/article/yazykovye-modeli (accessed: 24.05.2025). Russian.
2. Khokhlov AL, Kotlovskiy MYu, Pavlov AV, Potapov MP, Gabidullina LF, Tsybikova EB. Razvitiye neyrotekhnologiy: eticheskiye problemy i obshchestvennyye diskussii. Meditsinskaya etika. 2024; 1: 20–25. Russian.
3. Eticheskiye voprosy LLM-assistentov v klinike [Internet]. Meditsinskiye novosti i stat'i. 2025, may — [po sostoyaniyu na 18 iyunya 2025 goda]. Available from: https://nexusacademy.ru/tpost/95kvl3v0k1-eticheskie-voprosi-llm-assistentov-v-kli?ysclid=mc2hv983yr798737451 (accessed: 24.05.2025) Russian.
4. Beauchamp TL, Childress JF. Principles of biomedical ethics, fifth ed. New York: Oxford University Press. 2001; 454.
5. UN. Universal Declaration of Human Rights. New York, UN. [Internet]. Available from: https://www.un.org/en/about-us/universal-declaration-of-human-rights
6. Ayers JW. Internal Medicine Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. [Internet]. JAMA Internal Medicine Published online. 2023 April 28; 6. Available from: https://kstp.com/wp-content/uploads/2023/05/jamainternal_ayers_2023_oi_230030_1681999216.70842.pdf (accessed: 24.05.2025)
7. WHO. Ethics and governance of AI for health. [Internet]. 2023. Available from: https://betterghanadigest.com/2023/05/17/who-calls-for-safe-and-ethical-ai-for-health/ (accessed: 24.05.2025)
8. Yanshen Sun. Exploring the Deceptive Power of LLM-Generated Fake News: A Study of Real-World Detection Challenges [Internet]. 2023; 16. Available from: https://arxiv.org/html/2403.18249v1 (accessed: 24.05.2025)
9. Bloomberg. Economist Most Liveable Cities 2023 Ranking: Western Europe, Australia Top List. [Internet]. 2023. Available from: https://www.bloomberg.com/news/articles/2023-06-22/economist-most-liveable-cities-2023-ranking-western-europe-australia-top-list (accessed: 24.05.2025)
10. Domingo Stephen. Health Service Journal (HSJ). [Internet]. 2023. Available from: https://www.ccal.co.uk/post/three-takeaways-from-the-health-service-journal-hsj-digital-awards-2023 (accessed: 24.05.2025)

Литература

1. Косарев Е. А. Учебник по машинному обучению. Режим доступа: [Электронный ресурс] https://education.yandex.ru/handbook/ml/article/yazykovye-modeli (дата обращения: 24.05.2025).
2. Хохлов А. Л., Котловский М. Ю., Павлов А. В., Потапов М. П., Габидуллина Л. Ф., Цыбикова Э. Б. Развитие нейротехнологий: этические проблемы и общественные дискуссии. Медицинская этика. 2024; 1: 20–25.
3. Этические вопросы LLM-ассистентов в клинике [Интернет]. Медицинские новости и статьи. 2025, май — [по состоянию на 18 июня 2025 года]. Режим доступа: [Электронный ресурс] https://nexusacademy.ru/tpost/95kvl3v0k1-eticheskie-voprosi-llm-assistentov-v-kli?ysclid=mc2hv983yr798737451 (дата обращения: 24.05.2025)
4. Beauchamp TL, Childress JF. Principles of biomedical ethics, fifth ed. New York: Oxford University Press. 2001; 454.

5. UN. Universal Declaration of Human Rights. New York, UN. [Internet]. Available from: https://www.un.org/en/about-us/universal-declaration-of-human-rights
6. Ayers JW. Internal Medicine Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. [Internet]. JAMA Internal Medicine Published online. 2023 April 28; 6. Available from: https://kstp.com/wp-content/uploads/2023/05/jamainternal_ayers_2023_oi_230030_1681999216.70842.pdf (accessed: 24.05.2025)
7. WHO. Ethics and governance of AI for health. [Internet]. 2023. Available from: https://betterghanadigest.com/2023/05/17/who-calls-for-safe-and-ethical-ai-for-health/ (accessed: 24.05.2025)
8. Yanshen Sun. Exploring the Deceptive Power of LLM-Generated Fake News: A Study of Real-World Detection Challenges [Internet]. 2023; 16. Available from: https://arxiv.org/html/2403.18249v1 (accessed: 24.05.2025)
9. Bloomberg. Economist Most Liveable Cities 2023 Ranking: Western Europe, Australia Top List. [Internet]. 2023. Available from: https://www.bloomberg.com/news/articles/2023-06-22/economist-most-liveable-cities-2023-ranking-western-europe-australia-top-list
10. Доминго Стивен. Health Service Journal (HSJ). [Internet]. 2023. Available from: https://www.ccal.co.uk/post/three-takeaways-from-the-health-service-journal-hsj-digital-awards-2023 (accessed: 24.05.2025)