

AI ethics education as a tool to minimize risks in medicine

Nikoleta Leventi¹, Alexandrina Vodenicharova², Vidin Kirkov¹

Department of Social and Preventive Medicine and Disaster medicine, Faculty of Public Health „Prof. Dr. Tzekomir Vodenitcharov, PhD”, Medical University of Sofia, Sofia, Bulgaria¹

Department of Bioethics, Faculty of Public Health „Prof. Dr. Tzekomir Vodenitcharov, PhD”, Medical University of Sofia, Sofia, Bulgaria²

n.leventi@foz.mu-sofia.bg

Abstract: *The rapid integration of artificial intelligence (AI) into medical practice has created unprecedented opportunities for clinical decision support, diagnostics, and patient communication. At the same time, it has introduced new categories of risk, including hallucinated outputs, biased recommendations, privacy breaches, and overreliance by inexperienced practitioners. These risks are amplified in high-stakes environments such as medicine, where erroneous AI-generated information can directly impact patient safety. This paper argues that robust AI ethics education is an essential risks mitigation strategy for future healthcare professionals. Drawing on the results of a survey conducted among medical students, the study examines perceptions of AI use in medicine, the perceived importance of AI ethics education, and expectations regarding its content. Respondents showed clear optimism about the role of generative artificial intelligence (GenAI) in medicine, with strong majorities agreeing that it can positively change clinical practice, improve patient care, and find useful applications. At the same time, the high number of undecided responses (especially regarding patient-care quality) reflects ongoing uncertainty about reliability and safety. Students also strongly emphasized the need for structured AI ethics education, as most agreed that it raises awareness of ethical issues and should involve multidisciplinary expertise. Across all proposed topics, including data privacy, bias, explainability, safety, fairness, and autonomy, students rated the relevance of ethics-related content as high, indicating a clear expectation that medical curricula must prepare them to navigate the respective risks and limitations. By synthesizing these findings with current debates on AI governance and medical safety, the paper positions ethics-driven education as a foundational component of responsible AI deployment in healthcare. The paper argues for embedding AI ethics directly into medical curricula as a proactive risk-mitigation tool. The contribution of this work lies in demonstrating, through empirical evidence, that future clinicians recognize both the promise and the dangers of AI in medicine.*

Keywords: Artificial intelligence (AI), generative artificial intelligence (GenAI), medical ethics education, ethical risks in medicine, AI in healthcare.

1. Introduction

The rapid emergence of artificial intelligence has begun to reshape the landscape of modern medicine. Once limited to rule-based decision support systems and narrow diagnostic tools, AI in healthcare has evolved into a new generation of models capable of producing clinical text, summarizing patient histories, generating differential diagnoses, drafting treatment plans, and even simulating patient interactions [1]. These capabilities have made GenAI attractive to clinicians, educators, and healthcare institutions seeking to improve efficiency, reduce administrative burden, and expand access to medical expertise. As a result, AI systems are increasingly integrated into everyday clinical workflows, from triage chatbots and radiology reporting assistants to tools that support medical students in learning complex subjects [2]. This rapid adoption signals a profound shift: generative models are no longer peripheral technologies but emerging collaborators in the clinical environment [3].

Yet the same qualities that make AI powerful also introduce significant risks when applied in high-stakes medical contexts. Unlike traditional medical software, generative models do not guarantee factual accuracy, consistency, or transparency. They may produce hallucinated clinical information, fabricate references, or generate plausible-sounding but unsafe recommendations. Biases embedded in training data can lead to inequitable treatment suggestions, while privacy concerns arise when models inadvertently reproduce sensitive patient information [4]. Overreliance on AI outputs (particularly among inexperienced practitioners) can further amplify these risks, potentially undermining clinical judgment and patient safety. In medicine, where decisions carry immediate and sometimes irreversible consequences, such failures are not merely technical issues but ethical and professional challenges. The growing presence of AI in clinical settings therefore demands a deeper understanding of how future healthcare professionals perceive these risks and how prepared they feel to navigate them.

Despite the increasing visibility of AI in healthcare, a critical gap persists in medical education, namely, the absence of structured, focused training on AI ethics for minimizing risks.

Existing curricula often emphasize technological enthusiasm rather than critical evaluation, leaving students without the conceptual tools needed to assess AI-generated information, identify potential harms, or understand the boundaries of safe use. As AI becomes more embedded in clinical practice, this educational gap becomes increasingly consequential. Without targeted ethics education, future clinicians may lack the ability to recognize when AI outputs conflict with established medical standards, when biases may distort recommendations, or when human oversight is essential [5].

This paper addresses this gap by examining the role of AI ethics-driven education as a mitigation strategy for the safe use of AI in medicine. Drawing on survey data from medical students, it explores how emerging professionals perceive AI's benefits and risks, how they evaluate the importance of AI ethics education, and what content they believe such education should include. By situating these findings within broader debates on trustworthy AI and medical safety, the paper argues that ethics education is not an optional supplement but a foundational requirement for responsible AI integration in healthcare.

2. AI in Medicine: Opportunities and Risks

The emergence of artificial intelligence has opened a new chapter in the digital transformation of healthcare. Unlike earlier AI systems that relied on structured inputs and narrow predictive models, AI can produce coherent clinical text, synthesize complex information, and simulate reasoning processes in ways that resemble human communication. These capabilities have accelerated its adoption across medical education, clinical practice, and administrative workflows. Yet the same generative flexibility that makes these systems powerful also introduces new categories of risk that are particularly consequential in medicine. Understanding both the opportunities and the vulnerabilities is essential for developing responsible strategies for integration [6, 7].

2.1. Clinical applications

AI is increasingly used to support a wide range of clinical tasks. In documentation, models assist with drafting patient histories, summarizing clinical notes, and generating discharge instructions. Help reducing administrative burden and freeing clinicians to focus

on patient care. In diagnostics, AI tools can help generate differential diagnoses, interpret imaging findings, or highlight potential red flags in laboratory results. Medical education has also been transformed, as students use AI to explain complex concepts, simulate patient interviews, and practice clinical reasoning in low-risk environments. These applications demonstrate the potential of AI to enhance efficiency, support decision-making, and expand access to medical expertise, especially in resource-constrained settings [6, 7].

2.2. Documented risks

Despite these benefits, AI introduces significant risks that differ in nature and magnitude from those associated with traditional clinical technologies. One of the most widely recognized issues is hallucination, where the generation of plausible but factually incorrect or clinically unsafe information. In a medical context, hallucinations can lead to inaccurate diagnoses, inappropriate treatment suggestions, or fabricated references that mislead clinicians and students. Bias is another critical concern with models trained on unrepresentative or historically biased datasets that may produce inequitable recommendations, disproportionately affecting marginalized patient groups. Privacy risks arise when models reproduce sensitive patient information embedded in training data or when users input identifiable data into unsecured systems [8].

A further challenge is the lack of transparency of AI models. Their internal reasoning processes are not easily interpretable, making it difficult for clinicians to understand why a particular recommendation was generated or to assess its reliability. This lack of transparency complicates accountability, when an AI-generated suggestion contributes to patient harm.

3. Risks in Medicine and AI Ethics Education as a Mitigation Strategy

3.1. Why medicine is uniquely vulnerable

Medicine is particularly sensitive to the risks of AI for several reasons. First, clinical decisions often involve high stakes, where errors can lead to delayed diagnoses, inappropriate treatments, or irreversible harm. Unlike domains where AI mistakes may be inconvenient or costly, in healthcare they can be life-threatening. Second, medical information is inherently complex, context-dependent, and nuanced. AI models, which rely on statistical patterns rather than true understanding, may struggle with rare conditions or atypical specific health practices. Third, the clinician-patient relationship is built on trust, confidentiality, and professional accountability. Introducing AI into this relationship raises concerns about transparency, informed consent, and the preservation of human oversight [2].

Additionally, medical training emphasizes evidence-based practice, yet AI systems do not always align with established clinical guidelines. Their outputs may blend accurate information with inaccuracies that are difficult to detect, especially for learners. The rapid speed of AI development further complicates, leaving institutions without clear frameworks for safe deployment. Finally, the ethical diversity of global healthcare systems means that AI must operate within varied norms, and expectations [6].

3.2. The ethical foundations

The integration of AI into clinical environments introduces a set of risks that extend beyond technical malfunction and reach deeply into ethical, professional, and societal domains [2]. Because medical decisions directly affect human health and well-being, the ethical foundations guiding AI use in medicine must be exceptionally robust. Ethical principles traditionally applied to healthcare, such as beneficence, non-maleficence, autonomy, justice, and accountability, take on new significance when clinicians interact with generative models capable of producing authoritative-sounding but potentially flawed outputs. These

principles provide a framework for evaluating when and how AI systems should be used, what safeguards are necessary, and how responsibility should be distributed between human practitioners and algorithmic tools.

Beneficence and non-maleficence require that AI systems contribute to patient welfare while minimizing the risk of harm. Yet AI models may hallucinate clinical facts, misinterpret symptoms, or produce biased recommendations, challenging clinicians to critically assess AI-generated content rather than accept it at face value.

Autonomy demands that patients remain informed decision-makers, but opaque AI systems complicate transparency and informed consent. Justice requires equitable treatment across diverse populations, yet biases embedded in training data can lead to unequal outcomes. Accountability becomes especially complex when AI tools influence clinical decisions, and even determining who is responsible for errors, as the clinician, the institution, or the model's developers, remains an evolving ethical question. These principles highlight the need for clinicians to understand not only how AI works but also how its limitations intersect with core medical values.

AI ethics education emerges as a practical and necessary mitigation strategy precisely because it equips future healthcare professionals with the skills to navigate these risks. In this way, ethics education does more than raise awareness, it actively strengthens clinical safety. It empowers future clinicians to use generative AI responsibly, critically, and in alignment with the ethical commitments that define medical practice.

4. Survey Methodology

4.1. Participants

The cross-sectional study targeted first-year medical students enrolled in the Faculty of Medicine at the Medical University – Sofia. All students were invited to participate through an online announcement distributed between October 15 and November 30, 2025. Participation was entirely voluntary and anonymous, with no incentives offered and no academic consequences for declining. This approach ensured that students could respond freely and without pressure, thereby enhancing the reliability of the collected data. The sample represents individuals at the beginning of their medical training, a group for whom perceptions of AI and its ethical implications are particularly relevant as they form foundational professional attitudes.

4.2. Instruments

Data were collected using a structured web-based survey adapted from a previously validated instrument developed by Lukas Weidener and Michael Fischer [9, 10, 11]. The survey consisted of 53 items organized into six thematic sections, combining both closed-ended questions and Likert-scale statements.

- Part 1 gathered demographic and educational background information, enabling contextual interpretation of responses.
- Part 2 assessed students' prior exposure to AI-based (chat) applications, particularly conversational or genAI tools.
- Part 3 included statements evaluating students' perceptions of AI use in medicine, covering aspects such as usefulness, reliability, and potential risks.
- Part 4 focused on students' views regarding the teaching of AI in medical education.
- Part 5 explored attitudes toward AI ethics, including the importance of ethical principles and responsible use.

- Part 6 examined the perceived relevance of specific topics that could be included in an AI ethics curriculum.

Items in Parts 3 through 6 were rated on a 5- point Likert scale, allowing respondents to express varying degrees of agreement or perceived importance. The structure of the instrument ensured comprehensive coverage of both experiential and attitudinal dimensions related to AI in medicine.

4.3. Data Collection and Analysis

The survey was administered online to facilitate broad accessibility and to accommodate students' schedules. In total 144 students took participation, of 350 first academic year students of medicine.

5. Survey Results

The results reflect both enthusiasm and caution as students recognize the transformative potential of AI while simultaneously expressing concerns about reliability, safety, and the ethical challenges it introduces. The following subsections present the key findings across three major themes, namely, perceptions of AI in medicine, the perceived importance of AI ethics education, and the relevance of specific topics that should be included in such training.

5.1. Perceptions of AI in Medicine

The survey results show that first-year medical students hold a predominantly positive view of the potential role of artificial intelligence in healthcare. When asked whether AI can *positively change medicine, find useful applications, and improve the quality of patient care*, the majority of respondents expressed agreement.

Across the three statements, between 49 and 85 students selected "agree," and an additional 17 to 36 selected "strongly agree," indicating a strong overall endorsement of AI's transformative potential. Only a small minority expressed disagreement. A notable proportion remained undecided (particularly regarding AI's ability to improve patient care, where 60 students selected "undecided") suggesting that while enthusiasm is high, some students still lack sufficient exposure or confidence to fully evaluate AI's clinical impact. Overall, the data reflect a student cohort that is optimistic about AI's usefulness in medicine but still cautious about its reliability and real-world implications (see Table 1.).

Table 1: AI contribution, perceptions of AI in medicine.

Please indicate your degree of agreement with the following statements: AI in Medicine can...	[1. positively change medicine.]	[2. find useful applications in medicine.]	[5. improve the quality of patient care.]
1: I strongly disagree	4	2	3
2: I disagree	15	6	15
3: Undecided	36	15	60
4: I agree	66	85	49
5: I strongly agree	23	36	17

5.2. Importance of AI Ethics Education

Students demonstrated a clear recognition of the importance of AI ethics education within medical training. When evaluating whether AI ethics education *raises awareness of ethical issues in clinical practice and should involve experts from multiple disciplines*, the majority responded favorably. Between 59 and 64 students agreed with the statements, and an additional 30 to 42 strongly agreed, indicating broad support for integrating ethics instruction into the curriculum. Only a small number of respondents disagreed (15–14 students) or strongly disagreed (3–4 students), while a moderate group remained undecided (32–25 students). These results suggest that students not only value ethics education but also expect it to be interdisciplinary, reflecting the complex

nature of AI-related challenges in healthcare. The findings highlight a strong student-driven mandate for structured, comprehensive ethics training as a necessary component of preparing future clinicians for responsible AI use (see Table 2.).

Table 2: Importance of AI ethics education.

Please indicate your degree of agreement with the following statements: AI ethics education...	[4. contributes to raising awareness for ethical issues in clinical everyday life.]	[7. should involve experts from various fields (e.g., medicine, computer science, philosophy) to ensure a multidisciplinary perspective on AI ethics.]
1: I strongly disagree	3	4
2: I disagree	15	14
3: Undecided	32	25
4: I agree	64	59
5: I strongly agree	30	42

5.3. Expected Content of AI Ethics Education

When asked to assess the relevance of specific topics for inclusion in AI ethics education, students rated nearly all proposed themes as important. Topics such as *data privacy, bias, explainability, safety of AI-based applications, and fairness* received particularly high relevance scores, with the majority of responses falling into the "quite relevant" or "very relevant" categories. For example, 47 students rated *data privacy* as "very relevant," and 40 to 46 students rated *safety, fairness, and explainability* as "quite relevant." Even more conceptually complex topics (such as *informed consent* in the context of AI and *autonomy*) were viewed as relevant by most respondents, though with slightly more variability. Only a small minority rated any topic as "not relevant," and moderate relevance ratings were common across all categories, indicating thoughtful engagement rather than uniform endorsement. Overall, the results show that students expect AI ethics education to address a broad spectrum of issues, combining foundational ethical principles with practical considerations related to risk, transparency, and patient rights (see Table 3.).

6. Discussion with Interpretation of Findings

The findings of this study reveal a student cohort that is both optimistic about the potential of AI in medicine and acutely aware of the risks associated with its clinical use. The strong agreement that AI can positively transform healthcare, improve patient care, and find meaningful applications suggests that students recognize the growing relevance of AI-driven tools in future medical practice. At the same time, the substantial number of undecided responses (particularly regarding AI's impact on patient care) indicates uncertainty, likely coming from limited hands-on experience with AI systems or concerns about their reliability. This underscores the need for structured educational interventions that help students critically evaluate AI outputs rather than rely on intuition or external narratives.

The results also demonstrate a clear demand for AI ethics education. Students agreed that such education is essential for raising awareness of ethical issues and should involve multidisciplinary expertise. This aligns with broader discussions in the literature emphasizing that ethical challenges in AI cannot be addressed solely from a technical or medical perspective. The strong support of diverse curricular topics (including bias, data privacy, explainability, safety, fairness, and autonomy) further illustrates that students expect ethics education to be both conceptually rich and practically oriented.

Overall, the findings suggest that students are not passive recipients of technological change; they are actively seeking the knowledge and ethical grounding needed to navigate AI responsibly. This positions ethics-driven AI education as a crucial component of preparing future clinicians for safe and accountable AI issue

Table 3: AI contribution, perceptions of AI in medicine.

Please estimate the relevance of the following topics for education within your medical university studies:	[Informed Consent: Given the complexity of AI, it is questionable whether doctors will be able to understand the technology itself in the clinical	[Bias: The use of AI in medicine can lead to discrimination if the data used for training or programming the AI lack representativeness.]	[Data Privacy: As the use of AI in medicine involves highly sensitive patient data, security gaps or data misuse can have far-reaching consequences	[Explainability : Decisions made by AI-based applications cannot always be traced by the users due to the technical structure and complexity.]	[Safety of AI-based Applications: If AI-based applications are used for medical purposes, such as in diagnosis or treatment decision-making, faulty]	[Fairness: In addition to fairness in terms of equal treatment by the AI-based applications used (e.g., risk of bias and discrimination), access to]	[Autonomy: The use of AI in medicine can limit the autonomy of patients (e.g., regarding the use of AI in their own treatment) and doctors (e.g., i
1: Not relevant	9	9	4	5	3	6	9
2: Slightly relevant	18	28	19	18	11	21	26
3: Moderately relevant	44	38	27	50	38	46	37
4: Quite relevant	53	40	47	40	44	46	52
5: Very relevant	20	29	47	31	48	25	20

Conclusions

This study highlights the importance of integrating ethics-driven AI education into medical education as GenAI becomes increasingly embedded in clinical practice. The survey results show that first-year medical students view AI as a promising tool capable of improving efficiency and patient care, yet they remain cautious about its limitations. Their responses reflect a balanced understanding, with enthusiasm for innovation paired with recognition of the risks posed by hallucinations, bias, opacity, and overreliance. This dual perspective reinforces the need for educational frameworks that equip students to critically assess AI-generated information and apply it responsibly in clinical contexts.

A key insight from the study is the strong student support of AI ethics education. Participants not only affirmed its importance but also articulated clear expectations regarding its content. Topics such as data privacy, fairness, explainability, and safety were consistently rated as highly relevant, indicating that students understand the multifaceted nature of AI-related risks. Their emphasis on multidisciplinary teaching further suggests that effective AI ethics education must bridge medicine with computer science to provide a comprehensive foundation for responsible practice. The findings position ethics education not as an optional supplement but as a necessary mitigation strategy for reducing AI-related risks in medicine.

Future research should explore how AI ethics education can be effectively implemented and evaluated within medical curricula. Further studies could examine how students' perceptions evolve as they gain clinical experience and interact with AI tools in real-world settings. Additionally, comparative studies across institutions and countries would provide insight into how cultural, regulatory, and educational contexts shape attitudes toward AI and its ethical challenges.

Acknowledgement

This publication is part of a project on the topic "Trustworthy and Ethical AI in Healthcare – the Opinion of Future Medical Professionals", funded by the Medical Science Council of the Medical University - Sofia, GRANT-2025, Contract No. D-181/04.06.2025.

References

[1] Leventi, N. (2023). The Adoption of Information Technologies in Health Promotion. ISESIA, 18. Accessed at:

https://isesia.fmi.uni-sofia.bg/documents/ISESIA_2023_e-book.pdf#page=24

[2] Leventi, N. (2025). "The Digital Doctor: Navigating the Ethical Landscape for Artificial Intelligence in Healthcare", ISBN: 978-619-7452-35-8, Publisher: Faculty of Public Health "Prof. Dr. Tsekomir Vodenicharov, MD", in Bulgarian.

[3] Rabbani, S. A., El-Tanani, M., Sharma, S., Rabbani, S. S., El-Tanani, Y., Kumar, R., & Saini, M. (2025). Generative artificial intelligence in healthcare: applications, implementation challenges, and future directions. *BioMedInformatics*, 5(3), 37.

[4] Patias, I. (2026). GPT AS GPT Generative Pre-trained Transformer as General-Purpose Technology, ISBN:978-954-07-6293-7, St. Kliment Ohridski University Press, Sofia, Bulgaria

[5] Leventi, N., Vodenitcharova, A., & Popova, K. (2020). Ethical aspects of the use of innovative information technologies in clinical trials. *Proceedings of CBU in Medicine and Pharmacy...*, 1, 66.

[6] Vodenitcharova, A., Leventi, N., & Popova, K. (2022). Innovative Information Technologies in Medicine, the Ethical Aspects—Medical Students' Opinion. In *ISGT2022 conference*" (CEUR Volume 3191, ISSN: 16130073- <http://ceur-ws.org/Vol-3191/paper08.Pdf>).

[6] Ogut, E. (2025). AI in clinical medicine: challenges across diagnostic imaging, clinical decision support, surgery, pathology, and drug discovery. *Clinics and practice*, 15(9), 169.

[7] Patias, I., & Georgiev, V. (2020). The Use of Big Data in Medicine and Public Health Policy-Making: Opportunities and Challenges. In *Proceedings of the thirteenth International Conference on ISGT'2020*, Sofia, Bulgaria (pp. 7-13).

[8] Nastoska, A., Jancheska, B., Rizinski, M., & Trajanov, D. (2025). Evaluating trustworthiness in AI: Risks, metrics, and applications across industries. *Electronics*, 14(13), 2717.

[9] Weidener, L., & Fischer, M. (2024). Artificial intelligence in medicine: cross-sectional study among medical students on application, education, and ethical aspects. *JMIR medical education*, 10(1), e51247.

[10] Weidener, L., & Fischer, M. (2023). Teaching AI ethics in medical education: a scoping review of current literature and practices. *Perspectives on medical education*, 12(1), 399.

[11] Weidener, L., & Fischer, M. (2023). Artificial intelligence teaching as part of medical education: qualitative analysis of expert interviews. *JMIR Medical Education*, 9(1), e46428.