Bioethica

Vol 12, No 1 (2026)

BIOETHICA

The implementation of AI in medicine: ethical and legal considerations regarding the black box dilemma

Markus Meserth

doi: 10.12681/bioeth.45066

To cite this article:

Meserth, M. (2026). The implementation of AI in medicine: ethical and legal considerations regarding the black box dilemma. *Bioethica*, *12*(1), 19–32. https://doi.org/10.12681/bioeth.45066

ΠΡΩΤΟΤΥΠΗ ΕΡΓΑΣΙΑ

Η εφαρμογή της τεχνητής νοημοσύνης στην ιατρική: ηθικές και νομικές πτυχές σχετικά με το δίλημμα του «μαύρου κουτιού»

Markus Meserth[1,2,3]

[1]Έδρα Οικονομικού Ποινικού Δικαίου, Πανεπιστήμιο του Λιχτενστάιν

[2]Έδρα Ποινικού Δικαίου και Δικαίου Ανθρωπίνων Δικαιωμάτων, Πανεπιστήμιο της Βιέννης

[3]Ασκούμενος στην Εθνική Επιτροπή Βιοηθικής και Τεχνοηθικής, Ελλάδα

markus.b.meserth@uni-bayreuth.de

ΠΕΡΙΛΗΨΗ

Η τεχνητή νοημοσύνη (ΤΝ) γίνεται όλο και πιο διαδεδομένη σε διάφορους τομείς, υποσχόμενη να μεταμορφώσει μακροπρόθεσμα τον τρόπο ζωής μας. Η διερεύνηση των πιθανών εφαρμογών των συστημάτων ΤΝ βρίσκεται σε εξέλιξη και στον τομέα της υγειονομικής περίθαλψης. Αν και η νέα τεχνολογία έχει τη δυνατότητα να βελτιώσει την ιατρική περίθαλψη, είναι σημαντικό να μην ενθουσιαζόμαστε υπερβολικά εστιάζοντας αποκλειστικά στις ευκαιρίες που προσφέρει. Η χρήση αυτής της τεχνολογίας δεν είναι χωρίς κινδύνους. Το παρόν άρθρο παρέχει μια ολοκληρωμένη επισκόπηση των τεχνολογικών βασικών αρχών και εξετάζει τις ηθικές και νομικές επιπτώσεις της χρήσης της ΤΝ στην υγειονομική περίθαλψη. Επιπλέον, ασχολείται με το ζήτημα του ποιος μπορεί να θεωρηθεί υπεύθυνος για ζημίες που προκαλούνται από αυτοματοποιημένα συστήματα στον ευαίσθητο αυτό τομέα και πώς μπορεί να αντιμετωπιστεί αποτελεσματικά η διάχυση της ευθύνης. Το άρθρο επικεντρώνεται επίσης στο κεντρικό ζήτημα της εξηγήσιμης φύσης της ΤΝ και των λεγόμενων συστημάτων «μαύρου κουτιού». Η μονομερής απόδοση ευθύνης στον ιατρό ως τελικό υπεύθυνο λήψης αποφάσεων ενδέχεται να είναι ανεπαρκής, καθώς αυτό θα μπορούσε να υπονομεύσει την εμπιστοσύνη των ασθενών και των επαγγελματιών του ιατρικού κλάδου, εμποδίζοντας την υιοθέτηση συστημάτων ΤΝ. Αντ' αυτού, θα πρέπει να επιδιωχθεί μια συνεργατική προσέγγιση που θα ενσωματώνει όλους τους εμπλεκόμενους στη διαδικασία λήψης αποφάσεων, από τους προγραμματιστές συστημάτων έως τους ασθενείς.

Λέξεις-κλειδιά:
τεχνητή νοημοσύνη, εξηγησιμότητα, ευθύνη, ιατρική και υγειονομική περίθαλψη

BIOETHICA
BIOHΘIKA
Hellenic Republic
National Commission
for Bioethics & Technoethics
DEMOKRITOS

ORIGINAL ARTICLE

The implementation of AI in medicine: ethical and legal considerations regarding the black box dilemma

Markus Meserth[1,2,3]

[1] *Student Research Assistant, Professorship of Economic Criminal Law, Compliance and Digitalization at the University of Liechtenstein*

[2] *Research Assistant, Criminal Law and Human Rights Law, University Anniversary Fund of the City of Vienna, Austria*

[3] *Legal Trainee, National Commission for Bioethics and Technoethics, Greece*

markus.b.meserth@uni-bayreuth.de

ABSTRACT

Artificial Intelligence (AI) is becoming increasingly prevalent in various aspects of our lives, promising to transform the way we live in the long term. The exploration of potential applications of AI systems is also underway in the healthcare sector. While this new technology has the potential to improve medical care, it is important not to become overly enthusiastic and focus solely on the opportunities it offers. Nevertheless, using such technology is not without risks. This article provides a comprehensive overview of the technological fundamentals and considers the ethical and legal implications of using artificial intelligence in healthcare. Additionally, it deals with the question of who can be held liable for damages caused by automated systems in the sensitive field of medicine, and how the diffusion of responsibility can be effectively addressed. The article also focuses on the central issue of explainability in relation to AI and so-called "black box" systems. Attributing liability unilaterally to the physician as the final decision-maker may be insufficient, as this could undermine the trust of patients and medical professionals, hindering the adoption of AI systems. A collaborative approach that integrates everyone involved in the decision-making process – from system developers to patients – into an ongoing transformation process should be pursued instead.

Keywords:
artificial intelligence, explainability, liability, medicine and healthcare

1. Introduction

Recent and ongoing developments in artificial intelligence are driving significant technological advancements in nearly all areas of life. However, they also pose new ethical and legal challenges. This is particularly evident in the highly sensitive field of medicine. The use of AI in healthcare transforms the status quo, with experts predicting significant support and workload reduction for healthcare professionals, particularly in diagnostics and therapy.[1] In the best-case scenario, this could lead to an improvement in treatment quality. Therefore, the use of AI in medicine has the potential to benefit not only individual patients but also the whole healthcare system and requires more comprehensive consideration. However, this technical innovation also involves risks. Unlike tried-and-tested technologies, AI raises the issue of a "black box".[2] The results generated by autonomous systems are no longer fully traceable for either the end user or the original developer.[3] This is particularly problematic because the use of AI systems in the medical field can inevitably cause harm to patients. In such cases, the question arises of how to determine liability for damages caused by autonomous systems. For these systems to be used legally compliant, it is essential to clarify who is responsible for the decisions they make. Only legal certainty will enable this technology to be established and used beneficially in the long term. While the obvious choice from a practical perspective is to attribute the behaviour of an AI system to the person who makes the final decision, this approach often seems overly simplistic and short-sighted. Ultimately, it forces treating physicians into the role of symbolic liability servants, imposing the risks of the technology exclusively on them.[4] Therefore, it is advisable to address potential ethical and legal conflicts at an early stage, in order to prevent social concerns surrounding this new technology and, above all, to establish a practical framework for action. In general, artificial intelligence in medicine has the potential to enhance medical care. However, any legal uncertainties surrounding the use of AI represent an obstacle to its integration into everyday healthcare.[5] To provide clarity, this article offers an overview of the complex technological foundations and the ethical and legal considerations associated with the use of artificial intelligence in medicine. It outlines the conditions that must be met to realise the benefits of artificial intelligence while also respecting the rights and protection of patients.

2. Fundamentals of AI in medicine

The use of artificial intelligence in medicine is not entirely uncontroversial. Sceptics are increasingly voicing concerns about the dehumanisation of medical treatment.[6] However, the use of new technology also has advantages. It is expected to improve the overall healthcare system and to provide better medical care for patients. Due to the wide range of tasks that physicians must perform in a clinical context, as well as prevailing economic pressures in the medical sector, physicians often work to the point of exhaustion.[7]

1 *Schlicker N, Langer M, Hirsch M C. Wie vertrauenswürdig ist künstliche Intelligenz? Ein Modell für das Spannungsfeld zwischen Objektivität und Subjektivität. Innere Medizin 2023, 64:1051–1057, 1051.*

2 *Dettling H-U. Künstliche Intelligenz und digitale Unterstützung ärztlicher Entscheidungen in Diagnostik und Therapie. PharmR 2019, 41:633–642, 635.*

3 *Bathaee, Y. The Artificial Intelligence Black Box and the Failure of Intent and Causation. Harvard Journal of Law & Technology 2018, 31(2), 889–938, 891.*

4 *Beck S, Faber M, Gerndt S. Rechtliche Aspekte des Einsatzes von KI und Robotik in Medizin und Pflege. Ethik in der Medizin 2023, 35:247–263, 247.*

5 *Weicken E, et al. Schwerpunkt künstliche Intelligenz in der Medizin – rechtliche Aspekte bei der Nutzung großer Sprachmodelle im klinischen Alltag, Innere Medizin 2025, 66:436–441, 437.*

6 *Beck S, Faber M, Gerndt S. op.cit., 252.*

7 *Tretter M, Samhammer D, Dabrock P. Künstliche Intelligenz in der Medizin: Von Entlastungen und neuen Anforderungen im ärztlichen Handeln. Ethik in der Medizin 2024, 36:7–29, 11.*

Delegating tasks to automated systems could provide some relief. Furthermore, artificial intelligence opens up new possibilities in various medical subfields.

2.1 Fields of application for AI in medicine

There are numerous potential application areas of artificial intelligence in medicine, and AI systems are already being utilised in various ways within medical contexts. These systems support treating physicians in several ways, including image-based diagnostic systems.[8] Numerous AI systems have been successfully tested and implemented in fields such as dermatology[9], gastroenterology[10], and pathology[11]. Radiology is another promising field that was identified early on as an area of application for AI due to the large amounts of data that need to be collected and analysed.[12] Technical advancements in MRI scanning now enable the capture of more detailed images, facilitating earlier and more precise diagnoses of various types of cancer. However, as image quality improves, so does the effort required to analyse this image data. AI technology can evaluate vast amounts of data that physicians could never handle, making it particularly promising for pattern recognition in images. For instance, an AI model developed by Charité University Medicine in Berlin in June 2025 can recognise over 170 types of tumours and enables a precise diagnosis of brain tumours in 99.1 percent of cases.[13] The capability of such systems to perform at a level comparable to that of physicians for specific questions is already a reality. In some areas, these systems are even capable of surpassing the capabilities of physicians. A 2019 study found that malignant tumours were being diagnosed more reliably by AI-supported software than by its human counterpart.[14] Superhuman performance has also been observed in the detection of retinal diseases[15] and the prediction of heart attacks[16]. The economic impact of this new technology is as diverse as the potential applications in the healthcare sector. On the one hand, AI systems promise to reduce healthcare costs by facilitating diagnoses, preventing unnecessary treatment, and alleviating pressure on medical professionals. On the other hand, the high initial investment costs for purchasing and implementing the systems, combined with ongoing maintenance and training costs, present a significant financial challenge for medical institutions.[17] From an economic perspective, it is therefore necessary to carefully evaluate available resources and determine which AI systems can provide the most significant

8 *Dettling H-U. op.cit., 636.*

9 *Li Z, Koban KC, Schenck TL, Giunta RE, Li Q, Sun Y. Artificial Intelligence in dermatology image analysis. Current developments and future trends. J Clin Med 2022, 11(22), 6826.*

10 *Chen H, Sung JJY. Potentials of AI in medical image analysis in gastroenterology and hepatology. J Gastroenterol Hepatol 2021, 36(1):31–38, 31.*

11 *Meroueh C, Chen Z E. Artificial Intelligence in anatomical pathology. Building a strong foundation for precision medicine. Hum Pathol 2023, 132:31–38, 38.*

12 *Gore J C. Artificial intelligence in medical imaging. Magnetic Resonance Imaging 2020, 68:A1–A4.*

13 *Yuan, D, et al. CrossNN is an explainable framework for cross-platform DNA methylation-based classification of tumors. Nat Cancer (2025); see the press release:* https://www.charite.de/service/pressemitteilung/artikel/detail/tumordiagnostik_ki_modell_erkennt_mehr_als_170_krebsarten. *Accessed: 03. June 2025.*

14 *Liu, X. et al. A comparison of deep learning performance against healthcare professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. The Lancet Digital Health 2019, Volume 1, Issue 6, e271–e297.*

15 *De Fauw J. et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med 2018, 24(9):1342–1350, 1342.*

16 *Weng S F, Reps J, Kai J, Garibaldi J M, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PloS one 2017, 12(4):e0174944.*

17 *Elgin C Y, Elgin C. Ethical implications of AI-driven clinical decision support systems on healthcare resource allocation: a qualitative study of healthcare professionals' perspectives. BMC Medical Ethics 2024, 25:148.*

benefits in each clinical setting. Due to the wide range of applications of AI in medicine and its accuracy, this technology is not going to disappear from the medical context. Instead, its use is likely to intensify as a result of ongoing developments, making it essential to comprehensively examine the opportunities and risks of its increasing presence in everyday clinical practice.

2.2 The "black box" phenomenon

Although there are promising developments in the field of artificial intelligence, there are also risks associated with overly enthusiastic and unconsidered application. This is primarily due to the technological characteristics of AI-based systems. A fundamental distinction is often made between weak and strong AI.[18] While strong AI remains a vision for the future, the systems currently used in medicine are classified as weak AI. These systems can therefore only solve a specific problem. From a technical perspective, the various weak AI models differ in terms of their level of explainability. Deep neural networks, which are typically for AI systems, are so complex that the weighting of individual parameters for decision-making cannot be determined globally, i.e., for all possible predictions.[19] Consequently, many of these systems are referred to as "black boxes".[20] Although users can often access the inputs and outputs of the systems, they cannot see or understand the processes within them. One reason is that deep neural networks typically use models containing subsymbolic representations, meaning even the developers cannot understand their inner workings.[21] What has been described is often considered problematic, particularly in sensitive situations, and has negative implications for the implementation in clinical contexts. This characteristic makes it especially difficult to detect errors in the system's functioning, meaning that neither users nor those affected can trust such systems or make informed decisions based on their recommendations because they cannot understand what contributed to the decision-making process and to what extent. Since it is impossible to identify incorrect decisions or misleading recommendations, it can also be difficult or even impossible to assign responsibility appropriately and hold someone accountable for the resulting damages. Results and diagnoses must be traceable, not only in the event of damages, but in all medical decisions. In everyday clinical practice, physicians' actions are also controlled in this way.[22] However, if the decision-making process is no longer traceable, such control is not possible. Transparency and explainability are crucial principles for AI to gain the necessary trust of both patients and physicians. When AI systems are used as black boxes in critical areas such as diagnostics, it is essential that patients and physicians understand how the AI reaches its decisions.[23] Otherwise, the use of artificial intelligence in medicine would be difficult to justify ethically. In addition to monitoring the final results, the data fed into the system must also be checked beforehand. The data used to train AI systems must be accurate and unbiased. The more advanced the AI, the more complex the data it works with, and the more data is required. Despite the sensitivity of such data, data protection could easily be neglected.[24] Therefore, the use of AI involves considerable risks,

18 *Seng L. Maschinenethik und Künstliche Intelligenz. In: Bendel O (ed.). Handbuch Maschinenethik, Wiesbaden 2018, 185–205, 186.*

19 *Hacker P, Krestel R, Grundmann S, Naumann F. Explainable AI under contract and tort law: legal incentives and technical challenges. Artificial Intelligence and Law 2020, 28:415–439, 417.*

20 *Bathaee, Y. op.cit., passim.*

21 *Baum K, Mantel S, Schmidt E, Speith T. From Responsibility to Reason-Giving Explainable Artificial Intelligence, Philosophy & Technology 2022, 35:12, 2.*

22 *Lohmann A, Schömig A. „Digitale Transformation" im Krankenhaus. Gesellschaftliche und rechtliche Herausforderungen durch das Nebeneinander von Ärzten und Künstlicher Intelligenz. In: Beck S, Kusche, C, Valerius B (eds.). Digitalisierung, Automatisierung, KI und Recht, Baden-Baden 2020, 345–364, 352.*

23 *Pham T. Ethical and legal considerations in healthcare AI: innovation and policy for safe and fair use. R. Soc. Open Sci. 2025, 12: 241873, 4.*

24 *Lohmann A, Schömig A. op.cit., 353.*

particularly when processing sensitive health data, that is essential for designing, training and operating AI systems. It is crucial to monitor and ensure the accuracy and quality of the data.[25] Particular attention must therefore be paid to the use of sensitive personal data in accordance with data protection regulations.

3. Ethical considerations on the use of AI

In light of the aforementioned concerns regarding the use of artificial intelligence, as well as its wide range of applications in clinical contexts, it is crucial to ensure that new systems are implemented in an ethically acceptable manner. From an ethical perspective, a categorical ban on the use of AI in a medical context from the outset seems unjustifiable, given the enormous opportunities it presents. However, it is essential to acknowledge that AI systems in this area pose specific risks to vulnerable individuals.[26] Consequently, there exists a set of ethical principles that must be observed. The WHO has published a guide outlining these principles.[27] The central starting point is the autonomy of those affected, which must not be undermined by the use of AI. Furthermore, the systems should undergo regular review by supervisory bodies. Their results must be explainable and traceable. Ethical standards should effectively counteract fears of increasing dehumanisation and declining human interaction.

3.1 Challenges in real-world application

In order to realise the full potential of this technology, it is essential that automated systems used in the highly sensitive field of healthcare meet certain ethical standards as well as to acknowledge that artificial intelligence has certain limitations that need to be considered. This is due to the amount of data that is put into the system and the black box phenomenon mentioned earlier.

3.1.1 Bias

The development and training of AI algorithms typically involve the use of datasets from the healthcare sector. If these datasets contain underlying biases, the resulting AI system may perpetuate or exacerbate existing inequalities.[28] For instance, an AI system trained solely on data from one demographic group may perform poorly when applied to a different group. This can lead to unequal treatment or misdiagnoses. Furthermore, several studies have documented racial bias in AI decision support systems. For example, AI models developed to support dermatological diagnoses were predominantly trained on lighter skin tones, resulting in skin diseases being less accurately detected in patients with darker skin tones.[29] These inequalities reflect imbalances in the underlying training data, which often contains limited or poorly characterised data from different patient groups. Gender-specific biases have also emerged. For instance, some AI algorithms fail to consider gender-specific variations in symptoms.[30] This has the potential to result in erroneous diagnoses or delayed treatment of diseases that manifest distinctly in women compared to men. In addition, socioeconomic disparities can arise. Patients in low-income regions may have limited access to

25 *Frost Y, Steininger M, Vivekens S. Nutzen, Chancen, Risiken und Haftung bei der Verwendung von Künstlicher Intelligenz im Kontext der KI-Verordnung und KI-Haftungsrichtlinie. MPR 2024, 4–20, 6.*

26 *Tretter M, Samhammer D, Dabrock P. op.cit., 15.*

27 *See under:* https://www.who.int/publications/i/item/9789240029200. *Accessed: 03. June 2025.*

28 *Bekbolatova M, Mayer J, Ong CW, Toma M. Transformative Potential of AI in Healthcare: Definitions, Applications, and Navigating the Ethical Landscape and Public Perspectives. Healthcare 2024, 12, 125, 18.*

29 *Daneshjou R, Smith M P, Sun M D, Rotemberg V, Zou J. Lack of transparency and potential bias in artificial intelligence data sets and algorithms: A scoping review. JAMA Dermatol. 2021, 157, 1362–1369, 1362.*

30 *Parikh R B, Teeple S, Navathe A S. Addressing bias in artificial intelligence in health care. JAMA 2019, 322, 2377–2378, 2377.*

healthcare services, resulting in less representative data that algorithms may misinterpret.[31] Overall, these findings demonstrate that, if not carefully developed, validated and monitored across different patient groups, AI systems can unintentionally reinforce existing inequalities. To ensure the equitable development, efforts must be made to generate diverse datasets that represent people from all backgrounds.[32] Ultimately, AI models must be trained using balanced datasets that represent a broad spectrum of ethnicities, genders and socioeconomic groups in order to counteract any biases in the system at an early stage.

3.1.2 Explainability

Ensuring trustworthy AI in healthcare requires explainability. The black box problem, mentioned at the outset, is one of the most significant barriers to trust because it makes it difficult for physicians to understand the decision-making processes. The goal of explainable AI research is therefore to open up the black box, or at least make it more transparent.[33] Thus, the most important requirement is explainability. This would enable humans to understand the internal logic of AI systems and the reasons behind their outcomes. Explainability aims to provide users with understandable information about the model's internal workings and its results. These models should therefore be structured in a way that makes them understandable through a sequence of cognitively accessible deductive arguments.[34] Building trust in AI through explainability is particularly important for medical AI systems. This is because they are used in high-risk scenarios that significantly impact vulnerable individuals. When integrating AI systems into medical activities, it is crucial to consider the trustworthiness of these systems, as well as the trust that physicians and patients have towards them. Excessive or insufficient trust in AI can have fatal consequences. Too much trust can lead physicians to become overly reliant on this technology without exercising critical evaluation. At the same time, too little trust can result in physicians failing to capitalize on the benefits of this technology in their decision-making processes.[35] Explainability enables clinicians to evaluate the recommendations provided by a system based on their experience and clinical expertise, allowing them to make informed decisions. This helps them to make an informed decision about whether to rely on the system's recommendations, which can consequently strengthen their trust in the system. In cases where the model's recommendations do not align with the physician's expectations, explainability can verify the clinical relevance of the parameters considered by the system.[36] It is also of central importance to educate healthcare professionals about AI technologies and their interpretability. Training programmes that explain how AI systems work, how to interpret their results and how to communicate these findings to patients could further strengthen trust in AI-supported decisions.[37] The best possible and ethically acceptable socio-technical interaction can be ensured by explainability and comprehensively trained medical personnel.

3.2 Shared decision-making

In light of the considerations regarding explainability, the key question is how new AI systems can achieve the highest possible level of medical self-determination. Shared decision-making can provide a valuable basis.[38]

31 *Goktas P, Grzybowski A. Shaping the Future of Healthcare: Ethical Clinical Challenges and Pathways to Trustworthy AI. J. Clin. Med. 2025, 14, 1605, 10.*

32 *Bekbolatova M, Mayer J, Ong CW, Toma M. op.cit., 20.*

33 *Chau M, Rahman M G, Debnath T. From black box to clarity: Strategies for effective AI informed consent in healthcare. Artificial Intelligence In Medicine 2025, 167, 103169, 4.*

34 *Ferrario A, Loi M.How Explainability Contributes to Trust in AI. ACM Conference on Fairness, Accountability, and Transparency 2022 (FAccT '22), 1458. Available at SSRN:* https://ssrn.com/abstract=4020557. *Accessed: 03. June 2025.*

35 *Schlicker N, Langer M, Hirsch M C. op.cit., 1051, 1052.*

36 *Amann J, Blasimme A, Vayena E, Frey D, Madai V I. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020, 20:310, 5.*

37 *Chau M, Rahman M G, Debnath T. op.cit., 5.*

38 *Samhammer D. et al. Klinische Entscheidungsfindung mit*

In general decisions with significant moral implications should not be entirely delegated to systems.[39] This principle should also be applied in the field of medicine, where decisions have a significant impact on the quality of life of patients. When it comes to control, it is not enough to solely work with a system. It is also necessary to understand the consequences of using the particular system. However, actively deciding to use a system requires an understanding not only of how the system works, but also of the context in which it operates and the consequences of using it in that context. The central problem of attributing responsibility is whether it is appropriate to hold users who operate a system responsible, even if they do not understand it adequately.[40] Conversely, it is also problematic to hold the developers of the systems responsible, as they may have a more detailed technical understanding, yet have little influence on how the system is actually applied in a given context.[41] Generalisable statements about how AI systems should be interacted with in a medical context will be difficult to make. It depends on the type of interaction with the system and its design. The discussion about explainability ultimately stems from the technological dilemma of black box AI algorithms: the conflict between accuracy and opacity. Accurate AI algorithms are increasingly complex to understand, while understandable algorithms generally perform worse.[42] As greater autonomy in systems is accompanied by greater complexity, the question arises of how to ensure an understanding of, and profitable interaction with, AI systems. For this reason, it is important to consider explainability. Without explainability, a dilemma emerges in human-in-the-loop scenarios. Either the system's recommendation is pointless for the human decision-maker (when the human and the system agree, or when the decision-maker already has conclusive reasons), or the lack of explainability undermines understanding of the underlying decision, and thus the moral responsibility that humans are supposed to bear (when humans and systems disagree, but the human does not have conclusive reasons).[43] In summary, human decision-makers require explanations. Explainability cannot simply be provided by developers; it must be developed and maintained in dialogue with the relevant stakeholders.[44] However, decisions on whether to use the system and how to take its recommendations into account must be made by humans.[45] The requirement for a human to perform a final check before every medical decision is intended to minimise errors caused by AI and the associated risks. Otherwise it would be unclear who would be responsible in the event of damages.[46] In this context, therefore, physicians would not be responsible for the system's recommendation, but for their own decision.[47] AI systems should only make recommendations, leaving the final decision to a human decision-maker. Furthermore, it is crucial that AI systems are used only as supporting tools, in line with their strengths, and that physicians serve as the decisive controlling authority.

4. Legal considerations on the use of AI

Ultimately, the law must also respond appropriately

Künstlicher Intelligenz. Berlin 2020, 21–51, 21.

39 *Benzinger L, Ursin F, Balke WT, Kacprowski T, Salloch S. Should Artificial Intelligence be used to support clinical ethical decision-making? A systematic review of reasons. BMC Med Ethics 2023, 24(1):48, 2.*

40 *Samhammer D. et al. op.cit., 24.*

41 *Samhammer D. et al. ibidem.*

42 *Ursin F, Lindner F, Ropinski T, Salloch S, Timmermann C. Levels of explicability for medical artificial intelligence: What do we normatively need and what can we technically reach? Ethik in der Medizin 2023, 35:173–199, 175.*

43 *Baum K, Mantel S, Schmidt E, Speith T. op.cit., 17.*

44 *Asghari H, Birner N, Burchardt A, Dicks D, Faßbender J, Feldhus N et al. What to explain when explaining is difficult. An interdisciplinary primer on XAI and meaningful information in automated decision-making. Zenodo, 1.*

45 *Salloch S. Künstliche Intelligenz in der Ethik? Ethik in der Medizin 2023, 35:337–340, 339.*

46 *Lohmann A, Schömig A. op.cit., 355.*

47 *Schlör D, Hotho A. Verantwortungsvolle Empfehlungssysteme für die medizinische Diagnostik. In: Reder M, Koska C (eds.). Künstliche Intelligenz und ethische Verantwortung, Bielefeld 2024, 101–120, 112.*

to the risks described above. AI systems are difficult to predict and reconstruct ex post. Consequently, damages caused by interaction with an AI system are often challenging to prove, or at least to attribute beyond doubt to a specific individual. This creates the risk that injured parties will find themselves unprotected and not only affects those directly involved but also threatens to undermine societal acceptance of new technologies.

4.1 Legal review from the patient's perspective

It is essential to establish clear guidelines for responsibility and liability. This will guarantee patient protection and that the healthcare system remains accountable for the outcomes of AI-based treatment. Policymakers and judicial authorities must therefore develop frameworks that define the roles and responsibilities of AI developers and medical practitioners. These frameworks should ensure that physicians are adequately trained in using AI systems, that patients are informed about the decision-making process, and that responsibility is clearly defined in the event of errors or malpractice. Empirical evidence shows that patient involvement in the treatment process is crucial.[48] This includes providing them with comprehensive information that forms the basis for their consent and protects their right to self-determination.[49] How the information is provided and its specific contents depend on the circumstances of the individual case and the chosen treatment method. When using AI, one could argue that it is just another technological tool and that there is no need to provide separate information about it compared to other technologies.[50] However, the unique characteristics of decision support systems and the novelty of using AI in medical decision-making suggest that information about AI should be included in comprehensive explanations.[51] In order to maintain control, patients must develop a sufficient understanding of the technology and its impact on medical decisions. Once AI systems have become standard practice for a specific treatment, their use is covered by a physician's freedom to choose a treatment method. Until that point, however, other principles apply. Using cutting-edge methods is not in itself a treatment error. If it were, there would be no medical progress. However, there are special information obligations. This means that it must be explained that it is not yet a universally recognised standard method and that unknown risks cannot be ruled out. The black box phenomenon mentioned above should certainly be explained. From the patient's point of view, however, it is particularly important that those affected should have a party to whom they can address claims for appropriate compensation and that the enforcement of these claims should not be made more difficult as a result of the use of AI systems.[52] Open and transparent communication is essential, especially for patients with no prior knowledge of the technological principles or the strengths and weaknesses of the AI system in the specific medical application.

4.2 Legal review from the physician's perspective

From the perspective of the treating physician, legal uncertainties when dealing with AI should also be avoided. Not only does the law restrict AI technology in principle, but it also often provides incentives for its use.[53] To fully leverage the benefits of the new method without automatically making the physician the sole decision-maker and subjecting them to unjustified liability, it is desirable to establish a clearly defined framework for action.

4.2.1 Application

Neither medical law nor other standards prohibit the use of AI for treatment purposes. However, in

48 *Samhammer D. et al. op.cit., 32.*

49 *Samhammer D. et al. ibidem.*

50 *Samhammer D. et al. ibidem.*

51 *Eichelberger, J. Arzthaftung. In: Chibanguza K, Kuß C, Steege H (eds.). Künstliche Intelligenz: Recht und Praxis automatisierter und autonomer Systeme. Baden-Baden 2022, 655–674, 662.*

52 *Samhammer D. et al. op.cit., 34.*

53 *Hacker P, Krestel R, Grundmann S, Naumann F. op.cit., 436.*

principle, medical services are subject to the physician's prerogative and must be provided in person.[54] The delegation of medical services to non-medical auxiliary personnel, which is permitted under certain conditions, is not an option in the case of AI.[55] The reason is that AI lacks the capacity to hold a legal personality.[56] When it comes to medical malpractice law, the question of how AI will affect the liability of physicians is a major concern. Liability can arise if the physician fails to perform the treatment following the relevant medical standards or if the information provided by the physician is inadequate. Treatment must be performed in accordance with the scientific and proven medical standard at the time of treatment, i.e., lege artis.[57] In most cases, AI systems currently do not meet these standards.[58] However, physicians and patients can agree to deviate from the medical standard in cases involving cutting-edge methods, as mentioned earlier. As a general principle, though, the more significant the potential consequences of an incorrect decision, the more critically physicians should question the results.[59] Conversely, the liability risks arising from physicians deciding not to use AI should not be ignored. There is no general answer to the question of whether a physician can be held liable in such cases. Each case must be assessed on its circumstances. In the interest of medical progress, physicians may deviate from the standard of care under certain conditions, within the scope of their freedom to choose a course of treatment.[60] However, according to case law from the German Federal Court of Justice, physicians are only obliged to use new methods if they promise a lower risk or better chances of recovery for the patient than the previous standard, and if they are generally accepted within the scientific community.[61] Whether this applies in individual cases depends on various factors, including the quality of the AI system, its level of integration within medical practice, the relevant field of medicine, and the specifics of the clinical case. Therefore, physicians should familiarise themselves with technological advances and consider all available options within the scope of their therapeutic freedom.

4.2.2 Responsibility

The legal challenges are based on the inherent characteristics of AI systems, such as explainability. When using automated systems, it is essential to strike a balance between different interests to protect patients from a diffusion of responsibility, while also ensuring that individuals interacting with AI are not held unduly liable. The following legal considerations are guided by the above-described concept of shared decision-making. This concept is based on the idea that a significant degree of control is required for someone to be held responsible for decisions made when interacting with AI. Nevertheless, there is a risk that physicians will become liability servants due to their specific roles as treating individuals and supposed final decision-makers.[62] Conversely, overly restrictive regulations could prevent beneficial developments, particularly in AI research. Where AI systems make suggestions and a human selects from these or acts on them, the human remains fundamentally responsible for the decisions and actions taken. However, as we have seen, this can be problematic if it is unclear what basis the AI's

54 *Haftenberger A, Dierks C. Rechtliche Einordnung von künstlicher Intelligenz in der Inneren Medizin. Innere Medizin 2023, 64:1044–1050, 1047.*

55 *Haftenberger A, Dierks C. ibidem.*

56 *Riehm T. Nein zur ePerson - Gegen die Anerkennung einer digitalen Rechtspersönlichkeit. RDi 2020, 42–48, 48.*

57 *Beck S, Faber M, Gerndt S. op.cit., 253.*

58 *Beck S, Faber M, Gerndt S. ibidem.*

59 *Droste W. Intelligente Medizinprodukte: Verantwortlichkeiten des Herstellers und ärztliche Sorgfaltspflichten, MPR 2018, 109–115, 112.*

60 *Eichelberger, J. Arzthaftung im Kontext von KI und Robotik. In: Ruschemeier H, Steinrötter B. (eds.). Der Einsatz von KI & Robotik in der Medizin. Baden-Baden 2024, 97–114, 100.*

61 *Federal Court of Justice (BGH) ruling of September 22, 1987, ZR 238/86: The Federal Court of Justice also ruled that although patients cannot always expect the use of the latest therapies, physicians may be obliged to inform their patients about new methods under certain circumstances.*

62 *Samhammer D. et al. op.cit., 30.*

decision proposals are based on, how valid they are and what data the system was trained on.[63] Moreover, it is challenging to identify another responsible party because AI systems are difficult to predict ex ante and cannot be easily checked for errors ex post.[64] Therefore, solutions are needed to ensure an appropriate distribution of responsibility. The model should always form part of a more comprehensive assessment that incorporates medical expertise.[65] Physicians and other professionals must not be reduced to mere executors of AI systems. If there are sufficient, professionally justified reasons to believe that the model is incorrect in a particular case, its decision must be overruled.[66] Even if the physician's broader assessment is incorrect and the model prediction is correct, the physician is protected from claims of medical malpractice as long as their reasons for deviating from the model are justified by professional knowledge and conduct. Conversely, a failure to deviate from an incorrect model prediction would only constitute a breach of the standard of care if the reasons for deviation were evident to a specialist.[67] It is the physician who decides on a treatment course, not the AI. They are responsible for ensuring that their patient is treated correctly. Treating physicians may be liable for negligence if they make incorrect decisions based on the AI's results, for example, if they blindly trust the AI's recommendations. This violates the physician's duty of care and may preclude recourse against the AI product developer in certain cases. Finally, in some instances, the developer may also rely on the treating physician to question the results. Ultimately, however, this will depend on the individual case. Additionally, it also seems reasonable to involve developers and clinics more closely in liability issues.[68] A problem that exists in modern societies, not only in the context of AI, is that individual decisions are usually made in a collective context.[69] In the context of our considerations, this means that, in most cases, it is not the medical staff but rather the hospital management that decides on the use of AI, while the developers decide on the specific functioning of the systems.

5. Conclusion

The European Union is striving to create a legal framework for artificial intelligence. The European Union's AI Act came into force in August 2024. Although the AI regulation itself does not contain any liability provisions, as this article demonstrates, such provisions do exist. An essential component of addressing liability issues from a regulatory perspective was the proposed AI Liability Directive. The directive aimed to provide adequate protection against damages caused by AI systems. However, in February 2025, the European Commission withdrew the draft directive, citing a lack of consensus on key issues.[70] At the end of July 2025, the European Parliament published a comprehensive study that highlighted the risks of a regulatory vacuum: In the absence of a uniform European approach, Member States are increasingly likely to develop divergent regulatory frameworks for AI liability. Therefore, an efficient EU-wide regulation is the only way to avoid over-regulation. [71]

In summary, for AI systems to be used in a beneficial and legally compliant manner, it is essential to clarify what they are capable of doing. The AI systems in the medical field, for example, cannot and should not entirely replace physicians' decision-making. It

63 *Samhammer D. et al. idem, 31.*

64 *Samhammer D. et al. ibidem.*

65 *Froomkin M, Kerr I R, Pineau J. When AIs Outperform Doctors: Confronting the Challenges of a Tort-Induced Over-Reliance on Machine Learning. Arizona Law Review 2019, Vol. 61:33, 34.*

66 *Hacker P, Krestel R, Grundmann S, Naumann F. op.cit., 423.*

67 *Droste W. op.cit., 112.*

68 *Samhammer D. et al. op.cit., 47.*

69 *Baum K, Mantel S, Schmidt E, Speith T. op.cit., 9.*

70 *See under:* https://commission.europa.eu/document/download/7617998c-86e6-4a74-b33c-249e8a7938cd_en?filename=COM_2025_45_1_annexes_EN.pdf. *Accessed: 03. June 2025.*

71 *See under:* https://www.europarl.europa.eu/RegData/etudes/STUD/2025/776426/IUST_STU(2025)776426_EN.pdf. *Accessed: 24. July 2025.*

is not desirable to completely delegate decisions to AI systems, especially in healthcare settings. From a technical perspective, it is particularly essential to adopt a rather pragmatic approach. Close communication with developers is necessary, even during data collection, to generate accurate, robust, and unbiased models. Further development occurs in consultation with physicians and patients, based on these models. Ideally, the use of AI will enable those involved to make decisions based on more and better information than before. Nevertheless there needs to be an ongoing discussion about the use of AI in everyday clinical practice. Medical staff and patients must both be given the opportunity to share their concerns and experiences and to be heard. The role of the law is not to prohibit technological developments, but to ensure they are integrated into the legal framework in a wise manner. In addition to multidisciplinary stakeholder involvement, a trustworthy application requires clear ethical frameworks and evidence through clinical trials. This is a complex task that requires systematic evaluation, monitoring of transformation processes, training, and interdisciplinary discussions to ensure effective implementation.[72] Integrating AI training into medical curricula and providing regular continuing education opportunities is also becoming increasingly important. The proposal to instruct in-house experts, who work at the interface between information technology and everyday clinical practice, is also being discussed. One concrete example of this is the role of the Chief Digital Officer, whose current responsibilities cover the interface between medical practice and information technology, and which could be expanded to encompass AI.[73] The idea of having one individual in charge, similar to the role of a data protection officer, can result in a lack of accountability, with rule violations being blamed solely on the AI officer instead of addressing the underlying causes. To promote responsible and trustworthy use of AI, the multi-eye principle should therefore rather be adopted. Ultimately, the use of medical AI applications will lead to a shift in the skill set required of physicians. The focus will shift from purely competency-based skills to placing greater emphasis on communication skills. This is because the use of AI in medicine creates new challenges that primarily require communication-based solutions. In light of this paper, the belief that AI systems will soon replace radiologists is unfounded.[74] At most, these radiologists will be replaced by human physicians who know how to use and evaluate AI systems. Despite all technical advances, the most important relationship remains that between physician and patient, and ensuring a trusting and communicative relationship is crucial.

72 *Schlicker N, Langer M, Hirsch M C. op.cit., 1056.*

73 *Walchshofer M, Riedl R. Der Chief Digital Officer (CDO). Eine empirische Untersuchung. HMD Praxis der Wirtschaftsinformatik 2017, 54(3):324–337, 324.*

74 *Tretter M, Samhammer D, Dabrock P. op.cit., 12.*

Bibliography

Amann J, Blasimme A, Vayena E, Frey D, Madai VI. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020, 20:310.

Asghari H, Birner N, Burchardt A, Dicks D, Faßbender J, Feldhus N, Hewett F, Hofmann V, Kettemann MC, Schulz W, Simon J, Stolberg-Larsen J, Züger T. What to explain when explaining is difficult. An interdisciplinary primer on XAI and meaningful information in automated decision-making. Zenodo.

Bathaee, Y. The Artificial Intelligence Black Box and the Failure of Intent and Causation. Harvard Journal of Law & Technology 2018, 31(2), 889–938.

Baum K, Mantel S, Schmidt E, Speith T. From Responsibility to Reason-Giving Explainable Artificial Intelligence. Philosophy & Technology 2022, 35:12.

Beck S, Faber M, Gerndt S. Rechtliche Aspekte des Einsatzes von KI und Robotik in Medizin und Pflege. Ethik in der Medizin 2023, 35:247–263.

Bekbolatova M, Mayer J, Ong CW, Toma M. Transformative Potential of AI in Healthcare: Definitions, Applications, and Navigating the Ethical Landscape and Public Perspectives. Healthcare 2024, 12, 125.

Benzinger L, Ursin F, Balke WT, Kacprowski T, Salloch S. Should Artificial Intelligence be used to support clinical ethical decision-making? A systematic review of reasons. BMC Med Ethics 2023, 24(1):48.

Chau M, Rahman M G, Debnath T. From black box to clarity: Strategies for effective AI informed consent in healthcare. Artificial Intelligence In Medicine 2025, 167, 103169.

Chen H, Sung JJY. Potentials of AI in medical image analysis in gastroenterology and hepatology. J Gastroenterol Hepatol 2021, 36(1):31–38.

Daneshjou R, Smith M P, Sun MD, Rotemberg V, Zou J. Lack of transparency and potential bias in artificial intelligence data sets and algorithms: A scoping review. JAMA Dermatol. 2021, 157, 1362–1369.

De Fauw J. Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, Askham H, Glorot X, O'Donoghue B, Visentin D, van den Driessche G, Lakshminarayanan B, Meyer C, Mackinder F, Bouton S, Ayoub K, Chopra R, King D, Karthikesalingam A, Hughes CO, Raine R, Hughes J, Sim DA, Egan C, Tufail A, Montgomery H, Hassabis D, Rees G, Back T, Khaw PT, Suleyman M, Cornebise J, Keane PA, Ronneberger O. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med 2018, 24(9):1342–1350.

Dettling H-U. Künstliche Intelligenz und digitale Unterstützung ärztlicher Entscheidungen in Diagnostik und Therapie. PharmR 2019, 41:633–642.

Droste W. Intelligente Medizinprodukte: Verantwortlichkeiten des Herstellers und ärztliche Sorgfaltspflichten, MPR 2018, 109–115.

Eichelberger, J. Arzthaftung. In: Chibanguza K, Kuß C, Steege H (eds.). Künstliche Intelligenz: Recht und Praxis automatisierter und autonomer Systeme. Baden-Baden 2022, 655–674.

Elgin CY, Elgin C. Ethical implications of AI-driven clinical decision support systems on healthcare resource allocation: a qualitative study of healthcare professionals' perspectives. BMC Medical Ethics 2024, 25:148.

Ferrario A, Loi M. How Explainability Contributes to Trust in AI. ACM Conference on Fairness, Accountability, and Transparency 2022 (FAccT '22), 1458.

Froomkin M, Kerr IR, Pineau J. When AIs Outperform Doctors: Confronting the Challenges of a Tort-Induced Over-Reliance on Machine Learning. Arizona Law Review 2019, Vol. 61:33.

Frost Y, Steininger M, Vivekens S. Nutzen, Chancen, Risiken und Haftung bei der Verwendung von Künstlicher Intelligenz im Kontext der KI-Verordnung und KI-Haftungsrichtlinie. MPR 2024, 4–20.

Goktas P, Grzybowski A. Shaping the Future of Healthcare: Ethical Clinical Challenges and Pathways to Trustworthy AI. J. Clin. Med. 2025, 14, 1605.

Gore JC. Artificial intelligence in medical imaging. Magnetic Resonance Imaging 2020, 68:A1–A4.

Hacker P, Krestel R, Grundmann S, Naumann F. Explainable AI under contract and tort law: legal incentives and technical challenges. Artificial Intelligence and Law 2020, 28:415–439.

Haftenberger A, Dierks C. Rechtliche Einordnung von künstlicher Intelligenz in der Inneren Medizin. Innere Medizin 2023, 64:1044–1050.

Li Z, Koban KC, Schenck TL, Giunta RE, Li Q, Sun Y. Artificial Intelligence in dermatology image analysis. Current developments and future trends. J Clin Med 2022, 11(22).

Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, Mahendiran T, Moraes G, Shamdas M, Kern C, Ledsam JR, Schmid MK, Balaskas K, Topol EJ, Bachmann LM, Keane PA, Denniston AK. A comparison of deep learning performance against healthcare professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. The Lancet Digital Health 2019, Volume 1, Issue 6, e271–e297.

Lohmann A, Schömig A. „Digitale Transformation" im Krankenhaus. Gesellschaftliche und rechtliche Herausforderungen durch das Nebeneinander von Ärzten und Künstlicher Intelligenz. In: Beck S, Kusche C, Valerius B (eds.). Digitalisierung, Automatisierung, KI und Recht, Baden-Baden 2020, 345–364.

Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. JAMA 2019, 322, 2377–2378.

Pham T. Ethical and legal considerations in healthcare AI: innovation and policy for safe and fair use. R. Soc. Open Sci. 2025, 12: 241873.

Meroueh C, Chen ZE. Artificial Intelligence in anatomical pathology. Building a strong foundation for precision medicine. Hum Pathol 2023, 132:31–38.

Riehm T. Nein zur ePerson - Gegen die Anerkennung einer digitalen Rechtspersönlichkeit. RDi 2020, 42–48.

Salloch S. Künstliche Intelligenz in der Ethik? Ethik in der Medizin 2023, 35:337–340.

Samhammer D, Beck S, Budde K, Burchardt A, Faber M, Gerndt S, Möller S, Osmanodja B, Roller R, Dabrock P. Klinische Entscheidungsfindung mit Künstlicher Intelligenz. Berlin 2020, 21–51.

Schlicker N, Langer M, Hirsch MC. Wie vertrauenswürdig ist künstliche Intelligenz? Ein Modell für das Spannungsfeld zwischen Objektivität und Subjektivität. Innere Medizin 2023, 64:1051–1057.

Schlör D, Hotho A. Verantwortungsvolle Empfehlungssysteme für die medizinische Diagnostik. In: Reder M, Koska C (eds.). Künstliche Intelligenz und ethische Verantwortung, Bielefeld 2024, 101–120.

Seng L. Maschinenethik und Künstliche Intelligenz. In: Bendel O (ed.). Handbuch Maschinenethik, Wiesbaden 2018, 185–205.

Tretter M, Samhammer D, Dabrock P. Künstliche Intelligenz in der Medizin: Von Entlastungen und neuen Anforderungen im ärztlichen Handeln. Ethik in der Medizin 2024, 36:7–29.

Ursin F, Lindner F, Ropinski T, Salloch S, Timmermann C. Levels of explicability for medical artificial intelligence: What do we normatively need and what can we technically reach? Ethik in der Medizin 2023, 35:173–199.

Walchshofer M, Riedl R. Der Chief Digital Officer (CDO). Eine empirische Untersuchung. HMD Praxis der Wirtschaftsinformatik 2017, 54(3):324–337.

Weicken E, Mittermaier M, Hoeren T, Kliesch J, Wiegand T, Witzenrath M, Ballhausen M, Karagiannidis C, Sander LE, Gröschel MI. Schwerpunkt künstliche Intelligenz in der Medizin – rechtliche Aspekte bei der Nutzung großer Sprachmodelle im klinischen Alltag, Innere Medizin 2025, 66:436–441.

Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PloS one 2017, 12(4):e0174944.

Yuan D, Jugas R, Pokorna P, Sterba J, Slaby O, Schmid S, Siewert C, Osberg B, Capper D, Halldorsson S, Vik-Mo EO, Zeiner PS, Weber KJ, Harter PN, Thomas C, Albers A, Rechsteiner M, Reimann R, Appelt A, Schüller U, Jabareen N, Mackowiak S, Ishaque N, Eils R, Lukassen S, Euskirchen P. CrossNN is an explainable framework for cross-platform DNA methylation-based classification of tumors. Nat Cancer 2025, 6: 1283–1294.

Declaration of competing interest

The author declares that he has no competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.