

Levels of explicability for medical artificial intelligence: what do we normatively need and what can we technically reach?

Frank Ursin, Felix Lindner, Timo Ropinski, Sabine Salloch, Cristian
Timmermann

Angaben zur Veröffentlichung / Publication details:

Ursin, Frank, Felix Lindner, Timo Ropinski, Sabine Salloch, and Cristian Timmermann.
2023. "Levels of explicability for medical artificial intelligence: what do we normatively
need and what can we technically reach?" *Ethik in der Medizin* 35: 173–99.
<https://doi.org/10.1007/s00481-023-00761-x>.



Levels of explicability for medical artificial intelligence: What do we normatively need and what can we technically reach?

Frank Ursin · Felix Lindner · Timo Ropinski · Sabine Salloch · Cristian Timmermann

Received: 3 October 2022 / Accepted: 15 February 2023 / Published online: 3 April 2023
© The Author(s) 2023

Abstract

Definition of the problem The umbrella term “explicability” refers to the reduction of opacity of artificial intelligence (AI) systems. These efforts are challenging for medical AI applications because higher accuracy often comes at the cost of increased opacity. This entails ethical tensions because physicians and patients desire to trace how results are produced without compromising the performance of AI systems. The centrality of explicability within the informed consent process for medical AI systems compels an ethical reflection on the trade-offs. Which levels of explicability are needed to obtain informed consent when utilizing medical AI?

Arguments We proceed in five steps: First, we map the terms commonly associated with explicability as described in the ethics and computer science literature, i.e., disclosure, intelligibility, interpretability, and explainability. Second, we conduct a conceptual analysis of the ethical requirements for explicability when it comes to informed consent. Third, we distinguish hurdles for explicability in terms of epistemic and explanatory opacity. Fourth, this then allows to conclude the level of explicability physicians must reach and what patients can expect. In a final step, we show how the identified levels of explicability can technically be met from

✉ Dr. Frank Ursin · Prof. Dr. med. Dr. phil. Sabine Salloch
Institute for Ethics, History and Philosophy of Medicine, Hannover Medical School (MHH),
Carl-Neuberg-Str. 1, 30625 Hannover, Germany
E-Mail: ursin.frank@mh-hannover.de

Jun.-Prof. Dr. Felix Lindner
Institute for Artificial Intelligence, Ulm University, Ulm, Germany

Prof. Dr. Timo Ropinski
Visual Computing Group, Ulm University, Ulm, Germany

Dr. Cristian Timmermann
Ethics of Medicine, Medical Faculty, University of Augsburg, Augsburg, Germany

the perspective of computer science. Throughout our work, we take diagnostic AI systems in radiology as an example.

Conclusion We determined four levels of explicability that need to be distinguished for ethically defensible informed consent processes and showed how developers of medical AI can technically meet these requirements.

Keywords Explainability · Interpretability · Intelligibility · Transparency · Informed consent

Ebenen der Explizierbarkeit für medizinische künstliche Intelligenz: Was brauchen wir normativ und was können wir technisch erreichen?

Zusammenfassung

Definition des Problems Der Begriff Explizierbarkeit („explicability“) bezieht sich auf die Verringerung der Undurchsichtigkeit („opacity“) von künstlicher Intelligenz (KI). Diese Bemühungen werden als entscheidend für medizinische KI-Anwendungen angesehen, da in der Regel technisch bedingt Kompromisse zwischen Genauigkeit und Erklärbarkeit eingegangen werden müssen. Dies bringt ethische Fragen mit sich, da Ärzt:innen und Patient:innen nachvollziehen möchten, wie medizinische Entscheidungen zustande gekommen sind. Gleichzeitig soll die Genauigkeit der KI-Systeme möglichst hoch sein, ohne ethische Kompromisse eingehen zu müssen. Die zentrale Bedeutung der Explizierbarkeit veranlasst uns, den Prozess der informierten Einwilligung für medizinische KI-Systeme zu diskutieren. Welches Maß an Explizierbarkeit ist erforderlich, um eine ethisch gut begründete informierte Einwilligung beim Einsatz von KI-Systemen in der Medizin zu erreichen?

Argumente Wir gehen in fünf Schritten vor: Zunächst definieren wir die Begriffe, die in Ethik und Informatik üblicherweise mit Explizierbarkeit in Verbindung gebracht werden, d. h. Offenlegung („disclosure“), Verstehbarkeit („intelligibility“), Interpretierbarkeit („interpretability“) und Erklärbarkeit („explainability“). Zweitens führen wir eine konzeptuelle Analyse der ethischen Anforderungen an die Explizierbarkeit vor dem Hintergrund der informierten Einwilligung durch. Drittens unterscheiden wir Hindernisse für die Explizierbarkeit in Bezug auf epistemische und erklärende Undurchsichtigkeit. Daraus lässt sich ableiten, welche Ebenen der Explizierbarkeit Ärzt:innen anbieten müssen und was Patient:innen erwarten können. In einem letzten Schritt zeigen wir, wie die identifizierten Ebenen der Explizierbarkeit aus Sicht der Informatik technisch erfüllt werden können. In unserer Arbeit nehmen wir diagnostische KI-Systeme in der Radiologie als Anwendungsbeispiel.

Schlussfolgerung Es wurden vier Ebenen der Explizierbarkeit unterschieden, die für die Informationsvermittlung in ethisch vertretbaren Einwilligungsprozessen beim Einsatz medizinischer KI hilfreich sind. Diese Ebenen haben wir aus ethischer, regulatorischer und rechtlicher Analyse gewonnen. Außerdem haben wir gezeigt, wie Entwickler medizinischer KI diese Anforderungen technisch erfüllen können. Mit der Ausnahme, dass der Einsatz von KI überhaupt offengelegt werden sollte, ist die Bereitstellung von Verstehbarkeit, Interpretierbarkeit und Erklärbarkeit nicht in jedem Fall ethisch verpflichtend. In Bezug auf die Anwendbarkeit schlagen wir vor, individuell die Bedürfnisse der Patient:innen und die möglichen Folgen medi-

zinischer Entscheidungen zu berücksichtigen: Je invasiver und die Lebensqualität negativ beeinträchtigender und je weniger reversibel ein medizinischer Eingriff ist, desto mehr Ebenen der Explizierbarkeit sollten Ärzt:innen von sich aus anbieten.

Schlüsselwörter Erklärbarkeit · Interpretierbarkeit · Verstehbarkeit · Transparenz · Informierte Einwilligung

Introduction

The umbrella term explicability refers to increasing the understanding of black-box artificial intelligence (AI) systems (Robbins 2019). Reducing the opacity of black-box AI systems is crucial for medical AI applications because of the moral and professional responsibility of physicians to provide reasons for decisions (Swartout 1983). As such, commentators claim that medical AI systems “must have an explainable architecture, designed to align with human cognitive decision-making processes familiar to physicians, and directly tied to clinical evidence” (Char et al. 2020). This is because the output of medical AI systems provides reasons to justify further diagnostic and therapeutic decisions (Ferreira and Monteiro 2021). In contrast, other commentators state that accurate results for decisions in medicine are more important than the property to explain how results are produced (London 2019). Accordingly, a “defense of the black box” justifies its application in medicine if the “cost of a wrong answer is low relative to the value of a correct answer” (Holm 2019). This position prioritizes reducing harms from erroneous human decisions over having full understanding of how decisions were reached. In this paper, we mediate between these positions by distinguishing levels of explicability for obtaining valid informed consent to medical AI-aided procedures.

The discussion about explicability is rooted in a technological dilemma of black-box AI algorithms, namely the trade-off between accuracy and opacity (Reyes et al. 2020; London 2019). Accurate AI algorithms are increasingly difficult to comprehend, while comprehensible algorithms perform poorer. Understanding how automated decision support for doctors and patients came about is important when it comes to life-threatening surgical interventions or vigorous medications, as Smith (2018, pp. 149–150) implicates: “I don’t know why you are ill, but my computer says ‘take these pills’ [...] and recommends surgery.” This entails ethical tensions because in medicine both physicians and patients may have a strong interest in tracing how facts came about that have implications for further action without having to sacrifice performance.

The centrality of explicability in medical AI invites one to reflect on the ethical requirements for information disclosure to meet the demands of informed consent. Therefore, the aim of this paper is to determine the levels of explicability required for ethically defensible informed consent processes and how they can technically be met by developers of medical AI. To assist clinical decision-making, we will conclude by proposing four levels of explicability, i.e., disclosure, intelligibility, interpretability, and explainability, as a framework that will allow assessing the extensiveness of informing patients within the informed consent process. Our framework is normative,

so physicians can infer what information they should disclose to patients when they intend to use medical AI to assist in diagnostic or therapeutic decision-making.

Throughout this article, we take diagnostic systems in radiology as an example of clinical decision support systems (CDSS), because radiological AI-aided diagnostic systems represent the most advanced application of all medical AI developed to date, they are already commercially available and are in clinical use (American College of Radiology 2023; Muehlematter et al. 2021). AI-driven diagnostic systems are used for cancer screening (e.g., mammography, cf. Jairam and Ha 2022), neuroimaging (e.g., dementia, cf. Ursin et al. 2021a), ophthalmology (e.g., diabetic retinopathy, cf. Ursin et al. 2021b), and dermatology (e.g., skin cancer, cf. Beltrami et al. 2022). Our findings may have also relevance for nondiagnostic medical AI systems, such as those assisting in administrative processes or those integrated in medical devices like defibrillators adjusting automatically (Brown et al. 2022; Ranschaert et al. 2021).

Methodical procedure

We proceed in five steps: first, to derive the normative requirements for information content, we summarize the ethical demands for informed consent with a particular focus on the German healthcare system. We consider it an advantage to start from the acknowledged and established ethical and legal standards for informed consent in the German healthcare system, because this allows us to be as specific as possible while at the same time recognizing that we are not exhaustive globally. Although there is a plethora of legal, regulatory, social, and ethical issues regarding the question “what to tell the patient” when AI is involved (Cohen 2020; de Miguel et al. 2020; Mitchell and Ploem 2018), we are particularly interested in the normative sources of the extensiveness of information required for ethical decision-making.

Second, we map the conceptions that commonly fall under the umbrella term “explicitability” as described in the literature on ethical AI, policy frameworks, ethical guidelines, and even computer science. Our own contribution is to synthesize and apply the various notions of explicability in these different discourses to medical AI. The aim of the second step is to distinguish levels of explicability through a conceptual analysis. In recognizing that we cannot be exhaustive within the scope of this work to cover both the medical ethics and computer science literature, we have chosen a goal-oriented approach. Therefore, our knowledge base for the conceptual analysis rests on two targeted literature searches to identify key ideas in reviews. First, we selected reviews of guidelines for ethical AI (Fjeld et al. 2020; Hagendorff 2020; Floridi and Cowls 2019; Jobin et al. 2019). Second, we selected recent taxonomies of explainable AI (XAI) in computer science (Yang et al. 2022; Graziani et al. 2022; Barredo Arrieta et al. 2020; Miller 2019; Holzinger et al. 2019; Lipton 2018; Adadi and Berrada 2018; see appendix). Since there are only early attempts to harmonize the global taxonomy of XAI and in spite of a lacking consensus across disciplinary boundaries (Graziani et al. 2022; Miller 2019), we cover a representative scope of the relevant discussion in both fields of ethics and computer science. We attached our findings of the concept mapping as an appendix to this article

(see Table 4 in the appendix), to facilitate the replication of our interdisciplinary approach for other medical ethics assessments.

Third, we distinguish hurdles for explicability in terms of epistemic and explanatory opacity in building on the works of Ferretti et al. (2018) and Burrell (2016). This step aims at creating a list of criteria and questions for safeguarding the ethical utilization of medical AI against the background of the four levels of opacity by Ferretti et al. (2018). Fourth, we connect the normative requirements for informed consent with the levels of opacity and conclude which level of explicability physicians normatively must reach and what patients can expect. In a last step, we show how the identified levels of explicability can technically be met from the perspective of computer science. To this end, we discuss recent attempts of developing and deploying XAI in radiology.

Requirements for informed consent

We take the established framework of Faden et al. (1986, p. 274) as a starting point for distinguishing five elements of an ethically valid informed consent: information disclosure, comprehension, competence, voluntariness, and the consent itself. These five elements are acknowledged both globally (Eyal 2019) and nationally in Germany in the ethics literature (Becker 2019, pp. 16–26). Concerning the explicability of AI models as applied in various clinical fields all five elements can be reasonably discussed. However, issues of information disclosure and comprehension seem to gain particular importance with respect to medical AI due to the abovementioned black-box dilemma, whereas voluntariness, competence, and the consent itself are more likely to resemble “standard” settings of clinical care.

According to the traditional view of informed consent, information disclosure is closely linked to comprehension. Physicians are constantly asked to “tailor” their information in a way that is appropriate to the needs of the individual patient, e.g., by considering educational backgrounds, health literacy, or the patient’s current competence for intellectually grasping complex medical facts. Furthermore, it is well known that patients’ information needs differ and can be assessed systematically by applying standardized instruments (Christalle et al. 2019). In addition, various interventions for improving patients’ comprehension have been developed, including written, audiovisual and interactive digital materials (Glaser et al. 2020). However, empirical evidence still indicates that patients’ understanding of the information provided by physicians is often far from being optimal (Pietrzykowski and Smilowska 2021; Schenker et al. 2010). We conclude that each patient requires a specific content and quality of information to understand a medical procedure and consent to it.

Whereas, apparently, patients’ comprehension is closely linked to healthcare providers’ competence and practice in providing medical information. Recent ethical analyses highlight that comprehension and disclosure requirements rest on different normative sources: whereas disclosure aims at preventing illegitimate control over a person’s decision, the comprehension requirement focuses more on enabling

the decision-maker to decide for something concrete (and not for something else; Millum and Bromwich 2021).

Beside the ethical standards of informed consent, there are also legal obligations further specifying which concrete information need to be disclosed to patients. Meeting these requirements is a *sine qua non* for transforming an illegitimate act of violating the patient's bodily integrity into an act that is principally permissible for physicians (Becker 2019, p. 76). The German Civil Code (BGB) specifies that physicians must disclose information in a comprehensible manner on all circumstances that are essential for the treatment, in particular the diagnosis, the expected medical progression, and the therapy (BGB § 630c Abs. 2). Specifically, information on the nature, extent, execution, expected consequences and risks of the medical procedure as well as its necessity, urgency, suitability and prospects of success with regard to the diagnosis or therapy needs to be disclosed (BGB § 630e Abs. 1). The patient also has to be informed about alternatives and their different burdens, risks or chances of recovery (BGB § 630e Abs. 1). If a patient explicitly refuses being informed, he or she does not have to be informed (BGB § 630c Abs. 4 and 630e Abs. 3).

Against the background of such high requirements regarding the content and quality of information and the challenges which the practice of informed consent faces in clinical reality, the question arises how such requirements are met (or need to be further transformed) in light of a healthcare practice which is supported by AI-driven systems. In addition to issues of communication, the specific characteristics of medical AI necessitate developing new standards for disclosure and comprehension. For example, it matters whether the AI-driven system is already approved as a medical device so that physicians can trust its reliable functioning. If the system still has a novel character and clinical studies are missing, then higher demands are placed on physicians regarding both a critical benefit–risk assessment and informing patients about the details of the new treatment modality (ZEKO 2021).

The EU's General Data Protection Regulation (GDPR) adds to this already demanding patient–physician communication the “right not to be subject to a decision based solely on automated processing” (GDPR 2016, article 22). However, there is no legally binding “right to explanation” because this is specified only as a recital (GDPR 2016, recital 71), but data subjects must be provided with “meaningful information about the logic involved” (GDPR 2016, article 13.2.f; 14.2.g, and 15.1.h). We conclude that patients must be informed that a medical procedure is taking place at all (whether or not AI is involved) and physicians must disclose the circumstantial information of the medical procedure. However, the amount and quality of information regarding the nature, extent, execution, etc. of the medical procedure must be aligned to the comprehension capacity of the patient, unless he or she refuses to be informed.

Mapping concepts of explicability

Explicability is an ethical concept often used as an umbrella term that incorporates “the epistemological sense of ‘intelligibility’ (as an answer to the question ‘How does it work?’) and the ethical sense of ‘accountability’ (as an answer to the ques-

tion ‘Who is responsible for the way it works?’)’ (Floridi and Cowls 2019, p. 8). Although the same term is not used in computer science (see appendix), it has been introduced to the debate on ethical AI by Floridi et al. (2018). Notions of explicability can be found in high-level guidelines on ethical AI (AI4People in Floridi et al. 2018; Robbins 2019, p. 499), although its conceptual value has been contested (Ursin et al. 2022; Cortese et al. 2022; Wadden 2021; Krishnan 2020; Mittelstadt 2019; Robbins 2019).

Despite its frequent use, Jobin et al. (2019) found significant differences in the meaning and justification of terms related to explicability. In their scoping review of principles in 84 guidelines for ethical AI, they clustered eleven principles of which transparency was the most common, followed by explainability, explicability, understandability, interpretability, communication, disclosure, and showing. Floridi and Cowls (2019) synthesized six guidelines for ethical AI authored by high-profile initiatives with 49 principles in total into a five-principles approach. Hagendorff (2020) examined 22 guidelines for ethical AI and found 18 principles. He clustered the terms transparency and openness (16 mentions); explainability and interpretability (10 mentions); and openness, human oversight, control, and auditing (12 mentions). Fjeld et al. (2020) clustered under “transparency & explainability” the terms open source data and algorithms, notification when interacting with an AI, notification when AI makes a decision about an individual, regular reporting requirement, right to information, and open procurement (for governments).

There are three major shortcomings in using principles to guide ethical decisions. First, principles are vague and therefore difficult to interpret, second, principles can be in conflict with each other as the authors of the four *prima-facie* principles of biomedical ethics concede (Beauchamp and Childress 2019), and third, there is a lack of conceptual clarity because, e.g., explainability and transparency are often considered synonymous (Kazim and Koshiyama 2021; Robbins 2019). For example, Funer (2022) refrained from defining explicability, explainability, and transparency when discussing whether accuracy or comprehension should guide information disclosure in the clinical application of AI systems. Being aware of the conceptual ambiguities means that certain concepts apply in specific domains, so that, for example, explainability is used differently in ethics and in computer science (Powers and Ganascia 2020, pp. 29–33).

While these shortcomings can hinder the implementation of principles into practice (Morley et al. 2020), the medical domain has a long tradition of coherence approaches to translate “high-level commitments and principles into practical requirements and norms of good practice” (Mittelstadt 2019, p. 503). There are attempts to reconcile conceptual ambiguities between philosophy, computational disciplines, law, economics, and engineering (Mattingly-Jordan et al. 2022; Amann et al. 2020). To the best of our knowledge, the article by Miller (2019) is the most comprehensive attempt to bring together insights from social sciences and computer science for XAI but it does not explicitly tackle medicine. Therefore, we still lack interdisciplinary work on XAI that brings together medicine, ethics, and computer science concerning the information that must be provided when medical AI is utilized to justify diagnostic or treatment decisions—this is a gap we aim to help bridge.

Explicability is not considered a moral principle on its own, comparable to the other four principles of biomedical ethics (Cortese et al. 2022; Morley et al. 2020). It is linked to them by mostly instrumental chains to avoid harm and increase trust or performance (Ursin et al. 2022; McCoy et al. 2022). A system is explicable when it is explainable and interpretable, making it more transparent, therefore more accountable for human decision-making, human oversight, and justifiable decisions (Morley et al. 2020). The EU's Guideline on Trustworthy AI provides a definition for explicability (High Level Expert Group on Artificial Intelligence 2019, p. 13):

Explicability is crucial for building and maintaining users' trust in AI systems. This means that processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions—to the extent possible—explainable to those directly and indirectly affected. Without such information, a decision cannot be duly contested. [...] The degree to which explicability is needed is highly dependent on the context and the severity of the consequences if that output is erroneous or otherwise inaccurate.

Generally, we conclude that the inner workings of AI systems should be considered for decision making in general. Specifically, our analysis of the requirements for informed consent suggests that explicability should be considered for the clinical application of medical AI. Nevertheless, explicability faces four hurdles that need to be overcome.

Hurdles for explicability

Our contribution builds on a synthesis of Burrell's (2016) account on different forms of opacity with that of Ferretti et al. (2018) and its application to the informed consent process in the medical domain. The influential work of Ferretti et al. (2018) has already descriptively been adopted in other work on ethical issues of informed consent to AI applications (Goisau and Cano Abadfa 2022; Astromskè et al. 2021), but the perspective on AI developers by Burrell (2016) has been neglected so far. Our interdisciplinary approach acknowledges that the hurdles for explicability in the patient–physician encounter are intertwined with the interests of AI developers.

Burrell (2016, pp. 3–5) distinguishes between opacity as (1) intentional corporate secrecy, (2) technical illiteracy, and (3) the way algorithms operate at the scale of application. The cause for intentional secrecy may be a form of self-protection by companies intending to maintain “their trade secrets and competitive advantage” (Burrell 2016, p. 3). Low technical literacy, also to be understood as a general epistemic opacity, is common in patients as well as in physicians as reading and writing code remains inaccessible to the majority of the population (Burrell 2016, p. 4). Also developers are affected by opacity since algorithms and models are multicomponent systems built by teams; thus, programmers must also contend with a specific epistemic opacity, i.e., how algorithms operate at the scale of specific application.

Ferretti et al. (2018) distinguish between the (1) lack of disclosure, (2) general epistemic opacity, (3) specific epistemic opacity, as well as (4) explanatory opacity.

While Burrell (2016) used a conceptual approach to distinguish forms of opacity, Ferretti et al. (2018) examined the GDPR to identify rights for data subjects (Table 1). The most basic hurdle is when physicians themselves are not fully aware that they are working with an AI technology (Ferretti et al. 2018, pp. 326–327), e.g., when they assume they use a conventional picture archiving and communication system (PACS). Physicians might also be hesitant to disclose that an AI system is used at all, because of worries that patients have irrational fears towards AI systems and, therefore, chose out of paternalism not to disclose technical aspects.

The second hurdle is the general epistemic opacity of AI systems, i.e., that an agent does not understand the general principles of their design and functionality rather than the technical details (Ferretti et al. 2018, p. 328; Wachter et al. 2017a, p. 76; GDPR art. 13–15). This hurdle refers to the lack of meaningful background knowledge about the components of AI systems, the importance of training data, the data processing from inputs to outputs, and how rules are established for classifications by learning from examples (Ferretti et al. 2018, pp. 327–329). In accordance with the GDPR, these are the “logic, significance, envisaged consequences, and general functionality” (Wachter et al. 2017a, p. 78).

Like Ferretti et al. (2018, pp. 327–329) and Burrell (2016, p. 4), we see the need to draw an explicit distinction between how AI systems generally work (general epistemic opacity) and how a specific AI system works (specific epistemic opacity). To provide an illustrative example, it makes a difference whether (1) a patient might be satisfied by learning that AI systems can diagnose a disease from medical images, (2) a specific AI system is only able to discriminate between “likely has this disease” and “likely does not have this disease” (e.g., IDx-DR, cf. Ursin et al. 2021b), as well as a yet (3) hypothetical AI screening system for any disease in a given medical specialty. Similarly as with conventional medical devices, persons who are not familiar with the general functioning of that device must receive a general introduction to the technology to be able to consent to the procedure.

In light of its ethical, technical, and medical significance, the next hurdle for explicability is the specific epistemic opacity of a particular AI system considering its specific training data, the clinical relevance of specific inputs (feature relevance), the limitations of the spectrum of possible outputs, and the internal rules for data processing (Ferretti et al. 2018, pp. 327–329). Specific epistemic opacity relates “to the question of *how* an AI system provides a specific outcome” (Ferretti et al. 2018, p. 327). This is important because there might be biased training data and resulting rules, which lead to an unjustified discrimination of a patient or patient groups. In other words, specific epistemic opacity covers the hurdle for explicability when a particular AI system might not be the ideal choice to answer a clinical question or serve particular groups of patients.

The fourth hurdle is explanatory opacity that “relates to the question of *why* an AI system provides a specific outcome” (Ferretti et al. 2018, p. 329). This derives also from the “black box problem” that “occurs whenever the reasons why an AI decision-maker has arrived at its decision are not currently understandable to the patient or those involved in the patient’s care because the system itself is not understandable to either of these agents” (Wadden 2021, p. 4). To be able to provide effective,

Table 1 Levels of opacity based on Ferretti et al. (2018) and Burrell (2016)

Levels of opacity	Explanation	Causes	Corresponding level of explicability	Question	Solutions
1. Lack of disclosure	Data subjects are unaware of being subject to automated decision-making	Intentional corporate or state secrecy	Disclosure	Whether or not is an AI system applied?	Make code available for scrutiny, through regulatory means or algorithmic audit (carried out with or without corporate cooperation)
2. General episodic opacity	Originates from a general lack of understanding how AI systems learn, classify, and predict	Technical illiteracy	Intelligibility	How do AI systems generally work?	Promote computational thinking at all levels of education
3. Specific episodic opacity	Originates from a lack of understanding how a specific AI system learns, classifies, and predicts	Lack of understanding the rules an AI system follows	Interpretability	How does this specific AI system work?	Examine the capabilities and limitations of the specific AI system
4. Explanatory opacity	Lack of causal explanation between input and output	The way algorithms operate at the scale of application	Explainability	Why does the AI system provide a specific output?	Promote code audits and explainable AI techniques

AI artificial intelligence

efficient, and satisfactory justifications for individual decisions is crucial for further medical treatment (Holzinger et al. 2019).

Levels of explicability

The clinical application of medical AI requires informed consent as for any other clinical procedure (Neri et al. 2020; Brady and Neri 2020). The question of why AI systems should be treated differently than other technologies already used in patient care is caused by the inherent levels of opacity of AI. The difference between conventional radiological images and an AI system analyzing that data is that the interpretation of the imaging data by an AI system is an additional task that influences the interpretation of the physician analyzing that data. Therefore, we conclude that the four levels of opacity influence the patient–physician encounter and should be countered by corresponding levels of explicability (see next section, Table 2). The aspects of information disclosure, comprehension, competence, voluntariness, and the consent itself have to be secured by considering the technological particularities of such systems. The “black box problem” makes this process particularly challenging. We found that the need for four different levels of explicability not only derives from the requirements for informed consent and the norms identified in the mapping of ethical concepts, but also as a response to the above-mentioned levels of opacity.

There are various interpretations on the levels of explicability medical professionals should provide. Hitherto, Amann et al. (2020) distinguish only two levels: first, understanding how systems arrive at conclusions in general, and second, identifying features for an individual prediction. Floridi and Cowsls (2019) distinguish intelligibility and accountability. Jobin et al. (2019) claim that communication and disclosure are crucial to increase explicability, i.e., the fact that AI is used, the evidence-base for AI use, limitations of the AI algorithm, and auditability, thereby expanding the levels of explicability to four. The notion of four aspects is reasonable, although not tailored to medicine and healthcare, hence, being abstract and not rooted in the ethical requirements for informed consent.

To adapt the concept of explicability to medical practice, we propose four levels of explicability: disclosure, intelligibility, interpretability, and explainability (Table 2). We do so by synthesizing both the concept mapping of the ethical discourse as well as the prevalent concepts of XAI in computer science to mitigate the opacity hurdles for explicability (appendix 1). We enrich each level with an ethical guiding question and identify ethical implications specifically oriented to the patient–physician encounter. We proceed by discussing for each level which information should be given to overcome which hurdle, why this is ethically important and what the ethical implications are, respectively.

Disclosure

As AI supported decision-making is not the standard in medical practice and a technology still under early development, patients may have a general interest in knowing that a medical decision was influenced by AI (ZEKO 2021). The GDPR requires dis-

Table 2 Levels of explicability as applied for this work

Proposed ethical requirements for informed consent	Concepts in computer science	Ethical guiding questions	Type of hurdle (built on Ferretti et al. 2018 and Burrell 2016)	Ethical implications
1. Disclosure	(Use of AI is assumed)	Was an AI system used?	Lack of disclosure	Obligation not to deceive
2. Intelligibility	Intelligibility as decomposability	How do AI systems generally work (input, output, training data, parameter, calculation)?	General epistemic opacity (functioning of AI in general)	Obligation to communicate general risks deriving from the application of AI
3. Interpretability	Simulatability as global post hoc interpretability	How does that specific AI system work (input, output, training data, parameter, calculation)?	Specific epistemic opacity (functioning of a specific AI system)	Obligation to identify individual and group risks deriving from a specific AI system
4. Explainability	Explanation as local post hoc interpretability	Why did the AI system reach the particular decision that directly affects the patient?	Explanatory opacity	Obligation to be prepared to challenge or defend decisions (peer-disagreement)

AI artificial intelligence

closure for being subject to algorithmic decision-making (GDPR 2016, article 22). In Germany, it is legally required to inform about the nature of a diagnostic procedure (BGB § 630c Abs. 2).

From an ethical perspective, the clearest case requiring disclosure of having used medical AI is when asked by the patient. Physicians should generally not lie to patients—particularly on issues that are of their direct concern, such as the provenance of reasons backing a medical decision. Furthermore, to respect autonomy, physicians have an obligation not to deceive patients. Pretending that a medical decision was reached only through human wit would be a form of deception. Lastly, physicians should avoid withholding information patients are likely to want. If patients are generally interested whether new technologies have been used for their diagnosis or treatments, physicians would enter an obligation to share such information with the patient.

Therefore, to obtain an ethically valid informed consent, medical professionals need to disclose the application of medical AI within diagnosis or treatment. It is ethically not desirable to subject patients to algorithmic decision-making without their awareness.

Intelligibility

The aim of intelligibility is to counter general epistemic opacity, guided by the question “How do AI systems generally work?” This question can be split up into two further questions. The first question, “What are the parts of algorithmic models?” refers to the principle of decomposability, i.e., the “ability to explain each of the parts of a model (input, parameter and calculation)” (Barredo Arrieta et al. 2020, p. 88; Lipton 2018, p. 14). Decomposability is a condition for the second question, “How do AI systems learn from training data and generate an algorithmic models’ output?” referring to the ability to explain the general functioning of the technology. A general overview on how medical AI systems work may be of interest to those who have a moderate interest or skepticism towards new technologies, but have no particular worries about such systems. Informing about the general aspects of medical AI might be enough as far as physicians do not identify any factors that may position the patient as being at high risk.

To be intelligible, information needs to be provided on general risks and benefits inherited in the technology. Medical AI, especially in radiology, may be beneficial in having a high diagnostic accuracy, accelerating the radiological workflow or may potentially lead to less costly healthcare (Canadian Association of Radiologists 2019). Concerning risks, there are biases in the training data and once the model is trained there are algorithmic bias and automation bias as “the tendency for humans to favour machine-generated decisions” (Geis et al. 2019). In Germany, the obligation to communicate these general risks derives from the German Civil Code stating that specified information on the nature, extent, procedure, expected consequences, and risks must be disclosed (BGB § 630c Abs. 2).

Interpretability

Inspired by differentiations made in computer science (Molnar 2022, chapter 3) and radiology (Geis et al. 2019), we distinguish between explainability and interpretability. While explainability regards local post hoc explanations of individual predictions with the aid of explanatory methods, interpretability refers to the degree to which a human can predict a model's output by comprehending its inner workings. We refer to interpretability as a feature of a specific algorithm in contrast to intelligibility as a feature of the technology machine learning (ML) as a branch of AI. We draw this distinction because it makes a difference which specific algorithm is used for which purpose.

If there are 209 radiological AI systems commercially available (American College of Radiology 2023), why should we use exactly this or that one? The specific algorithm, let's say, can diagnose 15 different diseases, but the patient has the 16th so the algorithm is not suitable for broad screenings. If an algorithm's output is only binary, let's say, the patient has that disease or not, the algorithm should only be used if that is the diagnostic question. If an algorithm was only trained on data from Caucasian cis-males, then this is a limitation because the algorithm will not perform well on a diverse patient population (Obermeyer et al. 2019).

Therefore, there is an ethical obligation to inform oneself about vulnerable groups and conditions that may lead to inaccurate results in terms of accuracy, validity, uncertainty, and applicability as minimally acceptable criteria for interpretability (Arbelaez Ossa et al. 2022). Furthermore, once risks have been identified, there is the obligation to inform affected patients about individual or group risks. Lastly, the spectrum of possible outputs of a specific AI system must be disclosed. The German Civil Code demands that specific information on the suitability and prospects of success with regard to the diagnosis must be provided (BGB § 630c Abs. 2). In addition, well-informed patients may point out issues that physicians did not consider in their assessment.

We therefore agree with Ploug and Holm (2020) as well as Neri et al. (2020) that patients need to receive certain information to be able to contest AI driven diagnostics:

- Personal health data: information on the type and source of input data,
- Bias: information on (a) the character of the training data; (b) how training data were categorized by domain experts; (c) how the AI model was tested,
- Performance: information on (a) accuracy, specificity, and sensitivity; (b) how the performance was tested, and
- Decision: information on the (a) degree of human or algorithmic agency in making decisions; (b) that physicians are responsible for the final diagnosis.

Explainability

Explanations refer to causes and answer why-questions (Miller 2019, p. 6). Why-questions like "Why did that specific algorithm diagnose that condition?" are the most difficult to answer, because they are counterfactual, i.e., they exclude all other possible diagnoses. In medicine, this process is called differential diagnosis and it

resembles abductive reasoning, i.e., concluding that the most likely hypothesis is true because all other hypotheses cannot explain an event properly. This resembles the scientific process of falsification, i.e., the empirical falsifiability of hypotheses (that a patient may have one out of n possible conditions).

One has to differentiate what an explanation means for a physician and for a patient because they ask different questions due to their different interests. While the physician as a domain expert uses the causes of a condition as an explanation for a medical indication and thereby as a justification for further examination, diagnosis or treatment, a patient needs an explanation for understanding his or her condition, maybe to adjust lifestyle or increase compliance. In case of the patient's understanding, explanations are mostly intrinsic because they satisfy curiosity. In all other cases, explanations are instrumental, because they are means to achieve an end. Further instrumental reasons for explainability are examinations, to find meaning (reconcile contradictions and inconsistencies), to manage social interaction (share understanding), and persuasion or "assignments of blame" (Miller 2019, p. 9). We conclude that physicians and patients need different types (and thereby levels) of explanations.

Physicians as domain experts should have access to the explanation why an algorithm reached a certain decision because the algorithm's output justifies or contests their own decision (Henin and Le Métayer 2021). There are already preliminary proposals for situations where the AI's output conflicts with the physician's decision, but it is beyond the scope of this paper to consider all possible combinations of peer-disagreement between a physician, colleagues, and AI systems (Kempt and Nagel 2022). However, meeting a certain level of explicability has the advantage that physicians may be able to defend their position against peers for using or omitting the use of AI.

As medical technologies increase in complexity, we need to acknowledge that limitations of resources oblige us to set limits in how detailed a patient can expect to receive explanations. Beyond a certain level, we can even say that offering more details about the working of a particular medical technology becomes a supererogatory act. Under resources scarcity, spending excessive amount of time informing patients may also conflict with obligations towards other patients.

Applying the levels of explicability

The question remains how the levels of explicability can be applied within the information process to obtain informed consent. As the levels we propose represent a stepwise model of increased complexity, not every patient may request or need the highest level of explicability. One may consider stratified risk levels (unacceptable, high, and low or minimal risk) as the EU's AI regulatory framework does (European Commission 2021), but the risks have to be specific. In the literature, there is the proposal to concentrate on justifiability and contestability in high-stakes situations (Henin and Le Métayer 2021) or to provide minimally acceptable criteria for explainability (Arbelaez Ossa et al. 2022).

We propose two principles to translate theory into practice: first, tailoring the levels of explicability to patient needs and wishes and, second, tailoring the levels of explicability to the scope of a medical decision for a patient. Tailoring to patient needs and wishes honors the respect for autonomy from which both the “right to know” and the “right not to know” are derived. The “right not to know” is not an absolute right, but rather a right that must be “activated” (Andorno 2004), i.e., patients should be asked up to which level they wish to be informed. This requires at the very least that physicians disclose the fact that AI was used. We conceive that every possible combination can be met in reality: patients with high technical literacy and low health literacy may wish to know how the AI system works. While having general knowledge (therefore not “needing” intelligibility), they may request interpretability due to curiosity, but may not wish to get an explanation for the AI’s specific output. A patient with a low technical literacy and high health literacy may wish to get an explanation on the reasons backing a decision, but rejects elaborations on intelligibility and interpretability.

The second principle has been recently established by Funer (2022, p. 13) stating that “the greater the scope of a medical decision for a patient [is], the more normatively decisive the patient’s insight into the factors relevant to this decision and their interpretation in the context of the patient’s personal life.” While this principle is abstract, we suggest three criteria: *the level of invasiveness of the treatment, the reversibility of the treatment, and the risk of reducing the quality of life.* If an AI system diagnoses lung cancer based on a patient’s radiograph indicating surgery or chemotherapy, then this is invasive, not reversible and affects the quality of life. Therefore, in this case the highest level of explicability is appropriate. If an AI system automatically determines the age of persons by assessing radiographs, then the

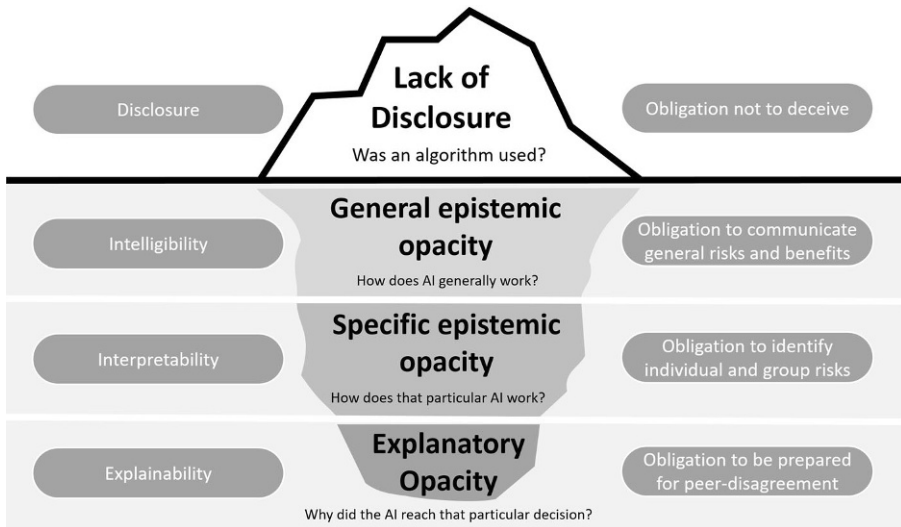


Fig. 1 Levels of explicability in relation to levels of opacity. How to read the iceberg shaped figure: start from the top and proceed to the bottom if the patient desires to get on the next level of explicability. AI artificial intelligence

lowest level of explicability is appropriate because this test does not entail invasive or not-reversible procedures and does not affect quality of life (Fig. 1).

XAI approaches in computer science

In a last step, we discuss whether and how the identified levels of explicability can technically be met from the perspective of computer science (Table 3). We rely on state-of-the-art XAI taxonomies (Graziani et al. 2022; Barredo Arrieta et al. 2020) and select five XAI methods specifically suited for visual data. To this end, we also discuss recent attempts of developing, deploying, and combining XAI in radiology (see section “Applying XAI in radiology”). The technical reasons for selecting these XAI methods are that their explanations can justify an AI systems output, serve control desires in terms of error identification, can improve the model itself, and enable to discover new facts and information (Adadi and Berrada 2018).

Table 3 distinguishes five main methods proposed for building XAI systems for visual data. One possibility is to train so-called inherently interpretable models instead of models as opaque as deep neural networks (DNNs). For instance, Li et al. (2012) train a K-nearest neighbor (KNN) model for image-based cancer prediction. Explaining how KNNs work in general is a simple task: They store each training image along with its classification label. To classify a new image as either showing cancer or not, the KNN looks up the K stored images that are closest to the new image. When the majority of these K stored images carry the “cancer” label, then the new image is also classified as cancer. In this sense, KNNs have a high degree of intelligibility. They are also interpretable because the internal structure of a trained KNN can be visualized by showing all the stored images, their mutual distances, and the assigned labels. Finally, to explain an individual prediction of a given image, the K-nearest neighbors of the given image and their labels can be shown: “This image has been classified as cancer because it is similar to these K images which also show cancer.” Unfortunately, inherently interpretable models such as KNNs achieve accuracy scores below DNNs.

Several methods have been developed that help understanding trained DNNs. Feature visualization (Nguyen et al. 2016) is a technique which is especially suitable for image classification with DNNs. A DNN consists of various layers of neurons

Table 3 Selection of five main methods for explainable artificial intelligence (AI) for visual data and how they meet the levels of explicability according to Graziani et al. (2022, table 6)

Method	Examples	Intelligibility	Interpretability	Explainability
Inherently interpretable models	K-nearest neighbor	x	x	x
Feature visualization	Preferred stimuli	x	x	x
Prototypes	MMD-Critic	–	x	x
Counterfactuals	Counterfactuals	–	–	x
Feature attribution	LIME, SHAP, saliency maps	–	–	x

LIME local interpretable model-agnostic explanations, *MMD* maximum mean discrepancy, *SHAP* Shapley additive explanations

that get activated by stimuli. The input image provides the stimulus for the first layer of neurons. If a neuron is sufficiently stimulated, it propagates a transformed stimulus further to the next layer of neurons. The idea behind feature visualization is to automatically generate images as input stimuli that maximize the activation of a neuron or a layer of neurons of interest. Some neurons may show high activity for images that contain a lot of edges, others may be more active for images with certain textures. This way, one can understand which parts of a given AI model respond to which parts of the input, and thus they contribute to a model's interpretability. The method may also be useful for how AI generally works (intelligibility). However, feature visualization does not provide an explanation of why a specific image was classified as, say, cancer.

Prototypes (Kim et al. 2016) can contribute to both interpretability and explainability. Prototypes are prototypical images from the training data set. By using the trained AI model to classify these prototypical images, one can get an overview of the model's behavior for a handful (or so) particularly representative images. An individual prediction can be explained by showing the closest prototype of the image input that receives the same prediction.

Often, the explainee is interested in knowing what parts of the image were particularly important for AI model's prediction. Counterfactuals (Wachter et al. 2017b) and feature attribution methods (Ribeiro et al. 2016; Lundberg and Lee 2017; Simonyan et al. 2013) can provide this information. Counterfactuals identify regions in the image that are important for the observed prediction, that is, when these regions were removed (greyed out), then the AI model would change its prediction. Feature attribution methods not only identify important regions but also assign importance values to regions in the image, viz., they can show which regions speak in favor of the AI model's prediction and which ones speak against it.

Applying XAI in radiology

Due to the medical relevance of XAI approaches, several review articles cover a wide range of explainability techniques in the medical domain (Yang et al. 2022; Graziani et al. 2022; Knapič et al. 2021; Adadi and Berrada 2020). In this section, we illustrate two such examples for radiology in detail. The first example has been chosen to show that even traditional saliency map techniques can still be improved. The second example demonstrates that a combination of XAI methods gains higher user satisfaction.

Saliency maps are likely the most commonly used XAI approach in medicine. By highlighting relevant regions, it is not only possible to provide reasoning behind an AI decision, but also to generate new knowledge, when the AI discovers features not represented in a labelled data set. Nevertheless, saliency techniques such as GradCAM (Selvaraju et al. 2016) often produce blurry highlights, which make localization difficult. To obtain more precise saliency maps, Major et al. (2020) have developed an approach which relies on image inpainting. During an inpainting process, they substitute healthy and unhealthy tissue. Based on the score difference of the images as well as saliency map quality, they can compute a saliency loss which is used in an iterative optimization to obtain sharper saliency maps. Their

results are demonstrated on mammograms and show a much more detailed saliency map when compared to state-of-the-art algorithms.

While saliency maps provide a good intuition about relevant regions within a radiological image, they lack the means to communicate via language. As highlighted image regions often require additional textual labels, or a text describing the particular medical context, natural language needs to be considered as an important part when considering XAI in radiology. Gale et al. (2018) were among the first to realize this demand by generating medical reports on radiological images. In their work, they propose an image-to-text model, which has been designed and trained to generate such reports for hip fractures from frontal pelvic x-rays. By combining a DenseNet with attention mechanisms, they are able to generate textual reports for x-ray images, whereby the report also contains details not provided through supervision. Thus, the authors were able to generate such sentences, which described the type of fracture and the location of the fracture with great accuracy, even outperforming the original reports. When confronting physicians with the outcomes of the system, they on average rated text alone (7.0) higher than saliency maps (4.4), while rating the combinations of the two best (8.8) on a 10-point Likert scale. This clearly shows that XAI should not only be visual, but should also consider other means of providing information.

Conclusion and future outlook

An ethically defensible information process when utilizing medical AI is possible through four levels of explicability as consecutive steps of escalation. Disclosure is the first condition to anticipate whether the patient desires further details. After the patient becomes aware of the intended use of medical AI, he or she is in a position to request further details or allow physicians to inform at their own discretion. Physicians should be able to offer to patients the further three levels, i.e., intelligibility, interpretability, and explainability to counter the epistemic and explanatory hurdles of medical AI. However, there is no ethical obligation to provide all three further levels in every case. In terms of applicability, we advise physicians to tailor the level of explicability to the needs of patients and the scope of the medical decision: the more invasive, the higher the effect upon quality of life, and the less reversible a medical decision is, the more levels of explicability should be provided.

We believe that our analysis of the explicatory hurdles of medical AI has implications not only for aligning information requirements in radiology in particular, but also for health care in general. We acknowledge that our analysis of the normative requirements for informing patients pose high stakes for the use of medical AI. This could lead to questioning the feasibility of these normative claims. However, instead of lowering the bar and reduce the explicatory burden physicians should bear, we rather suggest that the insights from medical AI ethics should be used to re-evaluate established medical practices and technologies. While we are increasingly learning about algorithmic biases of medical AI systems (Obermeyer et al. 2021, 2019), we also become aware that AI systems reproduce or exacerbate already existing biases, inequalities, and discriminations inherent in the training data (Ntoutsi et al. 2020).

This is not bad news, as we are now witnessing a window of opportunity to address the discriminatory effects that may also be prevalent in other sectors of medical practice today.

To inform about the quality of training data and its impact on marginalized population groups should not be a matter for medical AI alone, as increased awareness in fields such as dermatology is revealing. Darker skin types are underrepresented in cutaneous imaging data and models trained on these data perform poorly on patients with such skin tones (Kim et al. 2022). Discrimination does not only derive from the training data, but also from the classification system for skin types itself, the Fitzpatrick skin phototypes, because it does not capture variations in darker skin color and therefore restricts the range of options for people with darker skin. Ultimately, medical AI ethics can be an attention catalyst for confronting the biases long hidden in medical classification systems.

Appendix

Table 4 Concepts related to explicability in the domain of computer science. For the mapping, we used Yang et al. (2022); Graziani et al. (2022); Barredo Arrieta et al. (2020); Miller (2019); Holzinger et al. (2019); Lipton (2018)

Concept	Definition in computer science	Source
<i>Causability</i>	“The extent to which an explanation of a statement to a human expert achieves a specified level of causal understanding with effectiveness, efficiency and satisfaction in a specified context of use.”	Holzinger et al. (2019, p. 3)
<i>Comprehensibility</i>	“The ability of a learning algorithm to represent its learned knowledge in a human understandable fashion”	Barredo Arrieta et al. (2020, p. 84)
<i>Contestability</i>	“How the users can argue against a decision”	Yang et al. (2022, p. 31)
<i>Decomposability</i>	“Ability to explain each of the parts of a model (input, parameter and calculation), [... which] might empower the ability to understand, interpret or explain the behavior of a model”; equals intelligibility “Each part of the model—input, parameter, and calculation—admits an intuitive explanation. This accords with the property of intelligibility”	Barredo Arrieta et al. (2020, p. 88) Lipton (2018, p. 14)

Table 4 (Continued)

Concept	Definition in computer science	Source
<i>Explainability</i>	“To indicate with precision, to illustrate what features or high-level concepts were used by the ML system to generate predictions for one or multiple inputs.”	Graziani et al. (2022, table 3)
	“(Global) Explainable AI, also denoted as XAI, defines the branch of AI research that focuses on generating explanations for complex AI systems”; “feature attribution, feature visualization, concept attribution, surrogate, case-based and textual explanations”	Graziani et al. (2022, table 4)
	“Notion of explanation as an interface between humans and a decision maker that is, at the same time, both an accurate proxy of the decision maker and comprehensible to humans”	Barredo Arrieta et al. (2020, p. 85)
	“Explanation is post-hoc interpretability”	Miller (2019, p. 8); Lipton (2018)
	“Equat[ion of] ‘interpretability’ with ‘explainability’”	Miller (2019, p. 8)
	“Explanation for a wider range of users that how a decision has been drawn”	Yang et al. (2022, p. 31)
	“A collection of features of the interpretable domain, that have contributed to a given example to produce a decision”	Holzinger et al. (2019, p. 3)
<i>Simulatability</i>	“Ability of a model of being simulated or thought about strictly by a human”	Barredo Arrieta et al. (2020, p. 87)
	“A person can contemplate the entire model at once”	Lipton (2018, p. 13)
<i>Transparency</i>	“A transparent ML system has a non-opaque output-generation process where the role of the individual components, the learned paradigms, and the overall behavior of the model are known and can be simulated by a human user”	Graziani et al. (2022, table 3)
	“(Global) Transparency is used in AI to characterize those systems for which the role of internal components, paradigms and overall behaviour is known and can be simulated”	Graziani et al. (2022, table 4)
	“Level of accessibility to the data or model”	Yang et al. (2022, p. 31)
<i>Algorithmic transparency</i>	A model is transparent if it is understandable in three degrees of understandability: simulatable models, decomposable models and algorithmically transparent models	Barredo Arrieta et al. (2020, p. 85)
	“Ability of the user to understand the process followed by the model to produce any given output from its input data”	Barredo Arrieta et al. (2020, p. 88)
	“Transparency might apply at the level of the learning algorithm itself”	Lipton (2018, p. 14)

Table 4 (Continued)

Concept	Definition in computer science	Source
<i>Interpretability</i>	“To translate, expose, and comment on the generation process of one or multiple ML systems outcomes, making the overall process understandable by a human”	Graziani et al. (2022, table 3)
	“(Global) AI interpretability defines those AI systems for which it is possible to translate the working principles and outcomes in human-understandable language without affecting the validity of the system”	Graziani et al. (2022, table 4)
	“The ability to explain or to provide the meaning in understandable terms to a human”	Barredo Arrieta et al. (2020, p. 85)
	“Interpretability of a model [is] the degree to which an observer can understand the cause of a decision”	Miller (2019, p. 8)
	“Knowing how the AI technology functions”	Yang et al. (2022, p. 31)
	“A mapping of an abstract concept into a domain that the human expert can perceive and comprehend”	Holzinger et al. (2019, p. 3)
<i>Ante hoc interpretability</i>	“Ante-hoc systems are interpretable by design towards glass-box approaches [... by e.g.] linear regression, decision trees and fuzzy inference systems.”	Holzinger et al. (2019, p. 5)
<i>Post hoc interpretability</i>	“(Global) The AI system is neither inherently interpretable nor interpretable by-design, rather additional analyses are performed to generate explanations without re-training the model parameters”; “(i) feature attribution, (ii) feature visualization, (iii) concept attribution, (iv) surrogate explanations, (v) case-based explanations, and (vi) textual explanations.”	Graziani et al. (2022, table 4)
	“Extracting information from learned models [...] without sacrificing predictive performance.”	Lipton (2018, p. 15)
	“Posthoc systems aim to provide local explanations for a specific decision and make it reproducible on demand (instead of explaining the whole systems behavior).”	Holzinger et al. (2019, p. 5)
<i>Local interpretability</i>	“(Technical) Local interpretability is provided when interpretability analysis is performed on the system’s outcome for a single input”	Graziani et al. (2022, table 4)
<i>Global interpretability</i>	“(Technical) Global interpretability is provided when interpretability analysis is performed to explain the system behavior for a set of inputs corresponding to an entire class or multiple classes”	Graziani et al. (2022, table 4)
<i>Understandability</i>	Character “of a model to make a human understand its function—how the model works—without any need for explaining its internal structure or the algorithmic means by which the model processes data internally”	Barredo Arrieta et al. (2020, p. 84)
	“Understanding of the case to support a particular outcome”	Yang et al. (2022, p. 31)

AI artificial intelligence, ML machine learning, XAI explainable artificial intelligence

Author Contribution All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by all authors. The first draft of the manuscript was written by F. Ursin, section “Requirements for informed consent” was written by S. Salloch, section “XAI approaches in computer science” was written by F. Lindner, section “Applying XAI in radiology” was written by T. Ropinski, and all authors commented on previous versions of the manuscript. C. Timmermann substantially contributed to all sections of the manuscript. Figure 1 was drawn by F. Ursin. All authors read and approved the final manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest F. Ursin, F. Lindner, T. Ropinski, S. Salloch and C. Timmermann declare that they have no competing interests.

Ethical standards For this article no studies with human participants or animals were performed by any of the authors. All studies mentioned were in accordance with the ethical standards indicated in each case.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Adadi A, Berrada M (2018) Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* 6:52138–52160. <https://doi.org/10.1109/ACCESS.2018.2870052>
- Adadi A, Berrada M (2020) Explainable AI for healthcare: From black box to interpretable models. In: Bhateja V, Satapathy S, Satori H (eds) *Embedded systems and artificial intelligence*. *Advances in Intelligent Systems and Computing*, vol 1076. Springer, Singapore, pp 327–337 https://doi.org/10.1007/978-981-15-0947-6_31
- Amann J, Blasimme A, Vayena E, Frey D, Madai VI (2020) Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak* 20:310. <https://doi.org/10.1186/s12911-020-01332-6>
- American College of Radiology (2023) FDA cleared AI algorithms. <https://aicentral.acrdsi.org/>. Accessed 13 Jan 2023
- Andorno R (2004) The right not to know: an autonomy based approach. *J Med Ethics* 30(5):435–439. <https://doi.org/10.1136/jme.2002.001578>
- Arbelaez Ossa L, Starke G, Lorenzini G, Vogt JE, Shaw DM, Elger BS (2022) Re-focusing explainability in medicine. *Digit Health* 8:20552076221074488. <https://doi.org/10.1177/20552076221074488>
- Astromskė K, Peičius E, Astromskis P (2021) Ethical and legal challenges of informed consent applying artificial intelligence in medical diagnostic consultations. *AI Soc* 36:509–520. <https://doi.org/10.1007/s00146-020-01008-9>
- Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A et al (2020) Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion* 58:82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- Beauchamp TL, Childress JF (2019) *Principles of biomedical ethics*, 8th edn. Oxford University Press, New York
- Becker P (2019) *Patientenautonomie und informierte Einwilligung: Schlüssel und Barriere medizinischer Behandlungen*. J.B. Metzler, Berlin

- Beltrami EJ, Brown AC, Salmon PJM, Leffell DJ, Ko JM, Grant-Kels JM (2022) Artificial intelligence in the detection of skin cancer. *J Am Acad Dermatol* 87(6):1336–1342. <https://doi.org/10.1016/j.jaad.2022.08.028>
- Brady AP, Neri E (2020) Artificial Intelligence in radiology—Ethical considerations. *Diagnostics* 10(4):231. <https://doi.org/10.3390/diagnostics10040231>
- Brown G, Conway S, Ahmad M, Adegbe D, Patel N, Myneni V, Alradhawi M, Kumar N, Obaid DR, Pimenta D, Bray JJH (2022) Role of artificial intelligence in defibrillators: a narrative review. *Open Heart* 9(2):e1976. <https://doi.org/10.1136/openhrt-2022-001976>
- Burrell J (2016) How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data Soc.* <https://doi.org/10.1177/2053951715622512>
- Canadian Association of Radiologists (2019) White paper on ethical and legal issues related to artificial intelligence in radiology. *Can Assoc Radiol J* 70(2):107–118. <https://doi.org/10.1016/j.carj.2019.03.001>
- Char DS, Abràmoff MD, Feudtner C (2020) Identifying ethical considerations for machine learning health-care applications. *Am J Bioeth* 20(11):7–17. <https://doi.org/10.1080/15265161.2020.1819469>
- Christalle E, Zill JM, Frerichs W, Härter M, Nestoriuc Y, Dirmaier J et al (2019) Assessment of patient information needs: A systematic review of measures. *PLoS ONE* 14(1):e209165. <https://doi.org/10.1371/journal.pone.0209165>
- Cohen IG (2020) Informed consent and medical artificial intelligence: What to tell the patient? *Georget Law J* 108(6):1425–1469
- Cortese JFNB, Cozman FG, Lucca-Silveira MP, Bechara AF (2022) Should explainability be a fifth ethical principle in AI ethics? *Ai Ethics.* <https://doi.org/10.1007/s43681-022-00152-w>
- European Commission (2021) Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts. Brussels, 21.4.2021. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>. Accessed 13 Jan 2023
- Eyal N (2019) Informed Consent. In: Zalta EN (ed) *The Stanford encyclopedia of philosophy* (spring 2019 edition). <https://plato.stanford.edu/archives/spr2019/entries/informed-consent/>. Accessed 13 Jan 2023
- Faden RR, Beauchamp TL, King NMP (1986) *A history and theory of informed consent*. Oxford University Press, New York
- Ferreira JJ, Monteiro M (2021) The human-AI relationship in decision-making: AI explanation to support people on justifying their decisions <https://doi.org/10.48550/arXiv.2102.05460>
- Ferretti A, Schneider M, Blasimme A (2018) Machine learning in medicine. *Eur Data Prot Law Rev* 4(3):320–332. <https://doi.org/10.21552/edpl/2018/3/10>
- Fjeld J, Achten N, Hilligoss H, Nagy A, Srikumar M (2020) *Principled Artificial Intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI*. Berkman Klein Center Research Publication, vol 1. <https://doi.org/10.2139/ssrn.3518482>
- Floridi L, Cows J (2019) A unified framework of five principles for AI in society. *Harv Data Sci Rev.* <https://doi.org/10.1162/99608f92.8cd550d1>
- Floridi L, Cows J, Beltrametti M, Chatila R, Chazerand P, Dignum V et al (2018) AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds Mach* 28(4):689–707. <https://doi.org/10.1007/s11023-018-9482-5>
- Funer F (2022) Accuracy and interpretability: Struggling with the epistemic foundations of machine learning-generated medical information and their practical implications for the doctor-patient relationship. *Philos Technol* 35(1):5. <https://doi.org/10.1007/s13347-022-00505-7>
- Gale W, Oakden-Rayner L, Carneiro G, Bradley AP, Palmer LJ (2018) Producing radiologist-quality reports for interpretable artificial intelligence <https://doi.org/10.48550/arXiv.1806.00340>
- GDPR (2016) General Data Protection Regulation. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC. <https://eur-lex.europa.eu/eli/reg/2016/679/oj>. Accessed 13 Jan 2023
- Geis JR, Brady AP, Wu CC, Spencer J, Ranschaert E, Jaremko JL, Langer SG, Borondy Kitts A, Birch J, Shields WF et al (2019) Ethics of AI in radiology: Joint European and North American Multisociety statement. <https://www.acr.org/-/media/ACR/Files/Informatics/Ethics-of-AI-in-Radiology-European-and-North-American-Multisociety-Statement--6-13-2019.pdf>. Accessed 13 Jan 2023
- Glaser J, Nouri S, Fernandez A, Sudore RL, Schillinger D, Klein-Fedyshin M et al (2020) Interventions to improve patient comprehension in informed consent for medical and surgical procedures: An updated systematic review. *Med Decis Making* 40(2):119–143. <https://doi.org/10.1177/0272989X19896348>

- Goisau M, Cano Abadía M (2022) Ethics of AI in radiology: A review of ethical and societal implications. *Front Big Data* 5:850383. <https://doi.org/10.3389/fdata.2022.850383>
- Graziani M, Dutkiewicz L, Calvaresi D et al (2022) A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences. *Artif Intell Rev*. <https://doi.org/10.1007/s10462-022-10256-8>
- Hagendorff T (2020) The ethics of AI ethics: An evaluation of guidelines. *Minds Mach* 30:99–120. <https://doi.org/10.1007/s11023-020-09517-8>
- Henin C, Le Métayer D (2021) Beyond explainability: justifiability and contestability of algorithmic decision systems. *Ai Soc*. <https://doi.org/10.1007/s00146-021-01251-8>
- High Level Expert Group on Artificial Intelligence (2019) Ethics guidelines for trustworthy AI. <https://data.europa.eu/doi/10.2759/346720>. Accessed 13 Jan 2023
- Holm EA (2019) In defense of the black box. *Science* 364:26–27. <https://doi.org/10.1126/science.aax0162>
- Holzinger A, Langs G, Denk H, Zatloukal K, Müller H (2019) Causability and explainability of artificial intelligence in medicine. *Wires Data Min Knowl Discov*. <https://doi.org/10.1002/widm.1312>
- Jairam MP, Ha R (2022) A review of artificial intelligence in mammography. *Clin Imaging* 88:36–44. <https://doi.org/10.1016/j.clinimag.2022.05.005>
- Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. *Nat Mach Intell* 1(9):389–399. <https://doi.org/10.1038/S42256-019-0088-2>
- Kazim E, Koshiyama AS (2021) A high-level overview of AI ethics. *Patterns* 2(9):100314. <https://doi.org/10.1016/j.patter.2021.100314>
- Kempt H, Nagel SK (2022) Responsibility, second opinions and peer-disagreement: ethical and epistemological challenges of using AI in clinical diagnostic contexts. *J Med Ethics* 48(4):222–229. <https://doi.org/10.1136/medethics-2021-107440>
- Kim B, Khanna R, Koyejo OO (2016) Examples are not enough, learn to criticize! Criticism for interpretability. In: Lee D, Sugiyama M, Luxburg U, Guyon I, Garnett R (eds) *Advances in neural information processing systems*. Curran Associates, New York, pp 2288–2296 <https://doi.org/10.5555/3157096.3157352>
- Kim YH, Kobic A, Vidal NY (2022) Distribution of race and Fitzpatrick skin types in data sets for deep learning in dermatology: A systematic review. *J Am Acad Dermatol* 87(2):460–461. <https://doi.org/10.1016/j.jaad.2021.10.010>
- Knapič S, Malhi A, Saluja R, Främling K (2021) Explainable Artificial Intelligence for human decision support system in the medical domain. *Mach Learn Knowl Extr* 3(3):740–770. <https://doi.org/10.3390/make3030037>
- Krishnan M (2020) Against interpretability: a critical examination of the interpretability problem in machine learning. *Philos Technol* 33(3):487–502. <https://doi.org/10.1007/s13347-019-00372-9>
- Li C, Zhang S, Zhang H, Pang L, Lam K, Hui C, Zhang S (2012) Using the K-nearest neighbor algorithm for the classification of lymph node metastasis in gastric cancer. *Comput Math Methods Med* 2012:876545. <https://doi.org/10.1155/2012/876545>
- Lipton ZC (2018) The myths of model interpretability <https://doi.org/10.48550/arXiv.1606.03490>
- London AJ (2019) Artificial Intelligence and black-box medical decisions: Accuracy versus explainability. *Hastings Cent Rep* 49:15–21. <https://doi.org/10.1002/hast.973>
- Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: von Luxburg U, Guyon I, Bengio S, Wallach H, Fergus R (eds) *Advances in neural information processing systems*. Curran Associates, New York, pp 4768–4777 <https://doi.org/10.5555/3295222.3295230>
- Major D, Lenis D, Wimmer M, Sluiter G, Berg A, Bühler K (2020) Interpreting medical image classifiers by optimization based counterfactual impact analysis. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp 1096–1100 <https://doi.org/10.1109/ISBI45749.2020.9098681>
- Mattingly-Jordan S, Day R, Donaldson B, Gray P, Ingram LM (2022) Ethically aligned design. First edition glossary. https://standards.ieee.org/wp-content/uploads/import/documents/other/ead1e_glossary.pdf. Accessed 13 Jan 2023
- McCoy LG, Brenna CTA, Chen SS, Vold K, Das S (2022) Believing in black boxes: machine learning for healthcare does not need explainability to be evidence-based. *J Clin Epidemiol* 142:252–257. <https://doi.org/10.1016/j.jclinepi.2021.11.001>
- de Miguel I, Sanz B, Lazcoz G (2020) Machine learning in the EU health care context: exploring the ethical, legal and social issues. *Inf Commun Soc* 23(8):1139–1153
- Miller T (2019) Explanation in artificial intelligence: Insights from the social sciences. *Artif Intell* 267:1–38. <https://doi.org/10.1016/j.artint.2018.07.007>
- Millum J, Bromwich D (2021) Informed consent: What must be disclosed and what must be understood? *Am J Bioeth* 21(5):46–58. <https://doi.org/10.1080/15265161.2020.1863511>

- Mitchell C, Ploem C (2018) Legal challenges for the implementation of advanced clinical digital decision support systems in Europe. *J Clin Transl Res* 3(Suppl 3):424–430
- Mittelstadt B (2019) Principles alone cannot guarantee ethical AI. *Nat Mach Intell* 1(11):501–507. <https://doi.org/10.1038/s42256-019-0114-4>
- Molnar C (2022) Interpretable machine learning. A guide for making black box models explainable. <https://christophm.github.io/interpretable-ml-book/interpretability.html> Accessed 13 Jan 2023
- Morley J, Floridi L, Kinsey L, Elhalal A (2020) What to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Sci Eng Ethics* 26:2141–2168. <https://doi.org/10.1007/s11948-019-00165-5>
- Muehlematter UJ, Daniore P, Vokinger KN (2021) Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. *Lancet Digit Health*. [https://doi.org/10.1016/S2589-7500\(20\)30292-2](https://doi.org/10.1016/S2589-7500(20)30292-2)
- Neri E, Coppola F, Miele V et al (2020) Artificial intelligence: Who is responsible for the diagnosis? *Radiol Med* 125:517–521. <https://doi.org/10.1007/s11547-020-01135-9>
- Nguyen A, Dosovitskiy A, Yosinski J, Brox T, Clune J (2016) Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *Adv Neural Inf Process Syst* 29:3387–3395. <https://doi.org/10.48550/arXiv.1605.09304>
- Ntoutsis E, Fafalios P, Gadiraju U et al (2020) Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdiscip Rev Data Min Knowl Discov* 10(3):e1356. <https://doi.org/10.1002/widm.1356>
- Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366(6464):447–453. <https://doi.org/10.1126/science.aax2342>
- Obermeyer Z, Nissan R, Stern M, Eaneff S, Bembeneck EJ, Mullainathan S (2021) Algorithmic bias playbook. Center for Applied AI at Chicago Booth. https://www.ftc.gov/system/files/documents/public_events/1582978/algorithmic-bias-playbook.pdf. Accessed 13 Jan 2023
- Pietrzykowski T, Smilowska K (2021) The reality of informed consent: empirical studies on patient comprehension—systematic review. *Trials* 22(1):57. <https://doi.org/10.1186/s13063-020-04969-w>
- Ploug T, Holm S (2020) The four dimensions of contestable AI diagnostics—A patient-centric approach to explainable AI. *Artif Intell Med* 107:101901. <https://doi.org/10.1016/j.artmed.2020.101901>
- Powers TM, Ganascia JG (2020) The ethics of the ethics of AI. In: Dubber MD, Pasquale F, Das S (eds) *The Oxford handbook of ethics of AI*. Oxford University Press, New York, pp 27–51
- Ranschaert E, Topff L, Panykh O (2021) Optimization of radiology workflow with Artificial Intelligence. *Radiol Clin North Am* 59(6):955–966. <https://doi.org/10.1016/j.rcl.2021.06.006>
- Reyes M, Meier R, Pereira S, Silva CA, Dahlweid FM, von Tengge-Kobligk H et al (2020) On the interpretability of artificial intelligence in radiology: challenges and opportunities. *Radiol Artif Intell* 2:e190043. <https://doi.org/10.1148/ryai.2020190043>
- Ribeiro MT, Singh S, Guestrin C (2016) Why should I trust you? Explaining the predictions of any classifier. In: Krishnapuram B (ed) *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, New York, pp 1135–1144 <https://doi.org/10.1145/2939672.2939778>
- Robbins S (2019) A misdirected principle with a catch: Explicability for AI. *Minds Mach* 29:495–514. <https://doi.org/10.1007/s11023-019-09509-3>
- Schenker Y, Fernandez A, Sudore R, Schillinger D (2010) Interventions to improve patient comprehension in informed consent for medical and surgical procedures: a systematic review. *Med Decis Making* 31(1):151–173. <https://doi.org/10.1177/0272989X10364247>
- Selvaraju RR, Das A, Vedantam R, Cogswell M, Parikh D, Batra D (2016) Grad-CAM: Why did you say that? <https://doi.org/10.48550/arXiv.1611.07450>
- Simonyan K, Vedaldi A, Zisserman A (2013) Deep inside convolutional networks: Visualising image classification models and saliency maps <https://doi.org/10.48550/arXiv.1312.6034>
- Smith G (2018) *The AI delusion*. Oxford University Press, Oxford
- Swartout WR (1983) XPLAIN: a system for creating and explaining expert consulting programs. *Artif Intell* 21:285–325. [https://doi.org/10.1016/S0004-3702\(83\)80014-9](https://doi.org/10.1016/S0004-3702(83)80014-9)
- Ursin F, Timmermann C, Orzechowski M, Steger F (2021b) Diagnosing diabetic retinopathy with Artificial Intelligence: What information should be included to ensure ethical informed consent? *Front Med* 8(1108):695217. <https://doi.org/10.3389/fmed.2021.695217>
- Ursin F, Timmermann C, Steger F (2021a) Ethical implications of Alzheimer’s disease prediction in asymptomatic individuals through Artificial Intelligence. *Diagnostics* 11(3):440. <https://doi.org/10.3390/diagnostics11030440>

- Ursin F, Timmermann C, Steger F (2022) Explicability of artificial intelligence in radiology: Is a fifth bioethical principle conceptually necessary? *Bioethics* 36(2):143–153. <https://doi.org/10.1111/bioe.12918>
- Wachter S, Mittelstadt B, Floridi L (2017a) Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *Int Data Priv Law* 7(2):76–99. <https://doi.org/10.1093/idpl/ix005>
- Wachter S, Mittelstadt B, Russell C (2017b) Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv J Law Technol*. <https://doi.org/10.2139/ssrn.3063289>
- Wadden JJ (2021) Defining the undefinable: the black box problem in healthcare artificial intelligence. *J Med Ethics*. <https://doi.org/10.1136/medethics-2021-107529>
- Yang G, Ye Q, Xia J (2022) Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: a mini-review, two showcases and beyond. *Inf Fusion* 77:29–52. <https://doi.org/10.1016/j.inffus.2021.07.016>
- ZEKO (2021) Stellungnahme der Zentralen Kommission zur Wahrung ethischer Grundsätze in der Medizin und ihren Grenzgebieten (Zentrale Ethikkommission) bei der Bundesärztekammer „Entscheidungsunterstützung ärztlicher Tätigkeit durch Künstliche Intelligenz“. *Dtsch Arztebl* 118:33–34. https://doi.org/10.3238/arztebl.zeko_sn_cdss_2021