Ethics and Regulation of AI Systems in Medicine

The Example of Cancer Detection

Sebastian Bartsch, Marcus Düwell, Jan-Hendrik Schmidt, Alexander Benlian

Abstract *The integration of artificial intelligence (AI) systems into medical practice, specifically in cancer detection, presents unknown opportunities for better diagnoses and treatments for patients. However, with the integration of AI systems into a traditional relationship between healthcare professionals and patients, questions regarding accountability in this expanded relationship arise since traditional standards of medical law and medical ethics are addressed towards a healthcare professional. Against this backdrop, we investigate the necessary capacities to hold each involved party accountable (i.e., the healthcare professional, the patient, the developer, a regulatory oversight, and perhaps a clinical AI expert to support the healthcare professional). For this, we first explore the ethical and regulatory implications of employing AI systems in healthcare. We stress that the possibility of maintaining accountability is of central importance for the acceptability of the implementation of AI systems. As AI systems are often inscrutable and do not allow any party to explain and justify the behavior of the AI system, we examine whether and how explainable AI (XAI) methods can support each party with their accountability obligations. With our considerations, we propose a theoretical model for distributing accountability among each involved party and finally highlight the need for regulatory frameworks that can enable an ethically acceptable development and use of AI systems.*

1. Introduction

As the performance of AI systems continues to advance, they increasingly surpass human capacities in various domains. The more this will be the case, the more questions can be raised about the responsible development, deployment, and use of AI systems in practice. Given the variety of contexts in which AI systems are applied, each with its unique implications and challenges, we will limit our focus to the use of AI systems in medicine. In these contexts, the specific relationship between healthcare professionals and patients is of central importance, and therefore, the introduction of AI systems raises particular questions. For this reason, we also exclude possible AI systems designed for direct interaction with persons outside the regu-

lated healthcare environment. These systems would introduce distinct ethical questions regarding moral and legal responsibilities outside the highly regulated context of the healthcare system. In case AI systems take over tasks traditionally held by healthcare professionals, the challenge is that healthcare professionals do not only provide medical advice, but some sophisticated hermeneutical skills are required. By this we mean the following: Often patients are not fully able to understand the complicated medical information and they are often not aware of how they can respond to them. The healthcare professional needs skills to translate the information into a language a patient can understand and to interpret the wishes, expectations, hopes, and fears of the patient. Such hermeneutical skills are a necessary precondition for an accountable advice by the healthcare professional. This becomes more complicated if AIs are introduced in this relationship, and it raises complex ethical questions regarding standards of medical ethics that have been debated for decades (e.g., Fiske et al. 2019; Bringsjord 2008).

In this article, we focus on AI systems in the context of cancer detection. Focusing on this context is of high interest, as these AI systems are able to incorporate various data inputs to enable healthcare professionals to make more advanced and accurate diagnoses and treatments. However, our chosen example clearly presupposes that AI systems are solely used by the healthcare professional (perhaps with support from a clinical AI expert) and that AI systems are not directly used or interact with parties outside this healthcare setting. Additionally, we focus on such specific AI systems, as the example we have chosen is discussed in the overall context of this volume, focusing on normative questions that can be asked at a reflective level. Thus, we are not primarily evaluating whether some developments in AI systems are, as such, desirable, good, or risky. We rather focus on a question that necessarily has to be discussed before such an evaluation is possible, namely the question of what kind of capacities are required from the healthcare professionals as users of an AI system and how the relationship between the design of an AI system and the potential capacities of the healthcare professionals can raise questions of accountability. We are particularly interested in the question of how and to what extent healthcare professionals are capable of making informed judgments based on the advice AI systems provide. Suppose healthcare professionals are unable to understand how advice from AI systems is generated, and this raises the question to what extent the healthcare professional could be held accountable for the diagnosis and the treatment that is initiated on the basis of AI systems' advice. As a follow-up question, we will ask whether possible problems of accountability can be overcome by introducing clinical AI experts and oversight committees. Finally, we assess whether technical solutions like XAI are sufficiently capable of supporting healthcare professionals in their ability to deal with those systems in a responsible way.

We begin our article by (1) describing the technical functionality of AI systems in cancer detection, followed by (2) an exploration of the normative concerns these

systems raise. We then (3) evaluate the capacity of XAI to address the raised concerns and (4) discuss the need for additional regulatory measures to respond to those concerns.

2. AI Systems for Cancer Detection – How Does it Function?

Recent advancements in AI systems have shown significant promise in enhancing the accuracy and efficiency of cancer detection and diagnosis by supporting healthcare professionals (Zheng et al. 2023; Nassif et al. 2022). For instance, AI systems are capable of detecting various types of cancer, including skin cancer (e.g., Brancaccio et al. 2024), prostate cancer (e.g., Perincheri et al. 2021), gastrointestinal cancer (e.g., Suzuki et al. 2021), lung cancer (e.g., Ponnada and Srinivasu 2019), and, in particular, breast cancer (Zheng et al. 2023; Nassif et al. 2022). In general, such AI systems require cross-sectional images of suspected cancerous areas. In the example of detecting breast cancer, those images are often acquired through imaging modalities such as mammography for early-stage breast cancer detection (Zheng et al. 2023). Once captured, these images are processed and analyzed by the AI system, which has been trained to identify specific cancer types with high precision (Sechopoulos et al. 2021). Thus, the AI system is not omniscient and is only able to detect one type of cancer on which the AI system is trained and optimized (e.g., Russell and Norvig 2016).

These limitations of AI systems usually stem from the design and training process of the AI system (e.g., Russell and Norvig 2016). AI systems for cancer detection often employ a supervised learning approach, where the AI system is designed and trained for a single, specific diagnostic purpose (e.g., Shravya et al. 2019; Osareh and Shadgar 2010; Gupta and Gupta 2018). In this supervised learning training approach, developers train the AI system on a labelled dataset comprising both positive examples (i.e., images that contain the cancer to be identified) and negative examples (i.e., images that do not contain the cancer to be identified; e.g., Cunningham et al. 2008; Beeravolu et al. 2021). Accordingly, this training approach assigns the AI system a well-defined task and their desired solution of the task (often described as ground truth labels), enabling the AI system to iteratively adjust its parameters using statistical techniques until it generalizes the problem of detecting cancer with high accuracy. This iterative refinement process, referred to as *learning*, allows the AI system to progressively improve its predictive capabilities for the focused cancer type (Cunningham et al. 2008).

In designing AI systems for cancer detection, developers often employ neural networks that emulate the functioning of the human brain (Sechopoulos et al. 2021; Krogh 2008). A typical neural network comprises an input layer, one or more hidden layers, and an output layer, each containing nodes that perform specific mathemati-

cal operations when activated (Krogh 2008). These nodes on each layer are interconnected to form the network architecture, and developers can adjust the network's complexity by modifying the number of hidden layers and nodes to address various problems (Krogh 2008; Russell and Norvig 2016). Complex problems often require deeper neural networks with numerous layers and nodes, while simpler neural networks may suffice for less demanding problems (e.g., Hunter et al. 2012). However, as neural networks become more complex, they increasingly appear as *black boxes*, making it challenging for developers and users to understand how their AI systems derive specific outcomes (Asatiani et al. 2020; Berente et al. 2021). Through this resulting lack of transparency, developers may still evaluate an AI system's performance on test images by comparing the amount of correct or incorrect outcomes. However, they often cannot fully explain the reasoning behind the outcomes and are, thus, unable to fully understand the decision-making process of the AI system.

Since this lack of transparency is often critical in the productive use of AI systems, research on XAI seeks to address the challenge of understanding AI systems despite their inherent complexity (e.g., Adadi and Berrada 2018; Dwivedi et al. 2023). Notable advancements in XAI have led to the development of methods that analyze AI systems' behavior through external observation, enabling developers to interpret the system's functioning without direct insight into its internal architecture (e.g., Friedman 2001; Apley and Zhu 2020). By observing the AI systems' behavior, developers can derive aspects of the AI system's internal decision-making processes. Thus, developers approximate the AI systems' internal decision-making process but do not get a detailed explanation due to the limited interpretability of the system's inner architecture and mechanisms (Molnar 2019). Nevertheless, XAI methods provide developers transparency into their AI systems, with the aim of fostering user trust and acceptance of their AI systems (e.g., Dwivedi et al. 2023; Shin 2021). In this context, XAI methods can generally be divided into post-hoc and ante-hoc methods to gain transparency into the behavior of an AI system (Retzlaff et al. 2024).

Ante-hoc methods provide explanations inherently based on their algorithmic design and, thus, are limited to AI systems that use transparent algorithms such as linear regression and simple decision trees (Molnar 2019). On the other hand, post-hoc models aim to explain any designed and implemented AI systems by constructing additional surrounding models to explain either the AI system as a whole or specific instances within the dataset (Retzlaff et al. 2024). However, relying on a surrounding model implies that only an approximation of the AI system is explained, rather than the AI system itself. Consequently, post-hoc explanations may be inaccurate due to the inherent limitations of approximations, potentially undermining trust in their validity (Rudin 2019). Consequently, while ante-hoc models are generally more aligned with accurate explanations, this does not automatically mean that post-hoc approaches are inherently incapable of providing meaningful insights into the behavior of an AI system (e.g., Linardatos et al. 2020; Molnar 2019).

Classical representatives of post-hoc methods are partial dependency plots (PDP), permutation feature importance (PFI), local interpretable model-agnostic explanations (LIME), and influential instances (Molnar 2019). These representatives can further be divided into global-agnostic models, local-agnostic models, and instance-based methods to gain transparency into AI systems at different stages of the AI lifecycle (ibid.). In particular, global-agnostic models explain the AI system as a whole, while local-agnostic models explain the individual decisions of an AI system (Hariharan et al. 2023). Lastly, instance-based methods are used within the development of AI systems to gain information on how the AI system learns from the given dataset (Molnar 2019). Table **1** contains an overview of frequently used XAI methods and provides an explanation of each XAI method.

Table 1: Overview of frequently used XAI methods.

Category	**XAI Method (Examples)**	**Explanation of the XAI Method**	**Applicability**
Global-agnostic	Partial Dependency Plots (PDP)	PDPs illustrate the marginal effect of a feature (e.g., age, gender, or body weight) on the prediction of the AI system by averaging over all other features. Thus, PDP visualizes how changes in a given feature influence the prediction of the AI system. For example, a PDP can show how the probability of getting cancer changes when age increases by one year (Friedman 2001; Goldstein et al. 2015).	PDP is useful for understanding the importance of features during AI systems' development and validation.
Global-agnostic	Permutation Feature Importance (PFI)	PFI follows a similar principle compared to PDP but instead assesses feature importance by randomly shuffling values of a specific feature and measuring the resulting change in model performance. Thus, PFI quantifies how much a feature contributes to the prediction of the AI system. For example, while PDP reveals how the probability of getting cancer changes with age, PFI visualizes the relative importance of age for the prediction compared to other features, such as gender or body weight (Breiman 2001; Fisher et al. 2019).	PFI can be used to understand the dependency of the features of making predictions and can be used to validate and debug the AI system.

Category	XAI Method (Examples)	Explanation of the XAI Method	Applicability
Local-agnostic	Local Interpretable Model-agnostic Explanations (LIME)	LIME generates synthetic input data similar to a given instance and trains a transparent, interpretable AI system (e.g., linear regression) to approximate local decision boundaries. These boundaries provide users with insights into how the AI system behaves at a specific point. However, LIME's explanations are strictly local, meaning they apply only to a single instance rather than the entire AI system. For example, LIME can explain why an AI system predicts that a 58-year-old man weighing 65 kg has a low probability of developing cancer by showing that a normal body mass index decreases the risk of getting cancer. However, LIME cannot generally explain how the features of age, year, and body weight interact with each other and affect cancer (Ribeiro et al. 2016).	LIME can explain individual predictions to users after deploying the AI system.
Instance-based	Influential Instances	Influential instances are data points within the training set that significantly impact the predictions of the AI system. These instances can be problematic as their removal may cause substantial changes in AI systems' behavior. Identifying and managing influential instances enhances AI systems' robustness and performance while providing valuable insights into dataset composition (Molnar 2019).	Influential instances are useful during data preprocessing and debugging AI systems, as well as understanding the dataset.

Applying XAI methods in the context of cancer detection, healthcare professionals are not only informed of the presence and location of potential cancerous regions in the provided image but also receive insights into the AI system's decision-making process (Tosun et al. 2020). This additional information aims to enable healthcare professionals to integrate both the diagnostic findings and the AI systems' interpretative behavior into the patient's anamneses, guiding subsequent clinical decisions. Importantly, for these types of AI systems, even if they are used to support healthcare professionals with their clinical decisions, healthcare professionals remain the primary decision-makers, retaining the authority to accept or override the suggestions of the AI system (e.g., Tosun et al. 2020). Thus, this non-autonomous design of the AI systems aims to ensure that the system serves as a decision support system, with ultimate decision-making power residing with the healthcare professionals,

who are accountable for determining the course of treatment (Cooper et al. 2022). This specific setting leads, however, to the question of whether the healthcare professional is able to deal with these AI systems in a responsible way – a question that shall now be addressed.

3. Ethical Questions

In the first instance, it seems evident that the healthcare professional is still in charge of the situation. The healthcare professional bears accountability vis-à-vis the patient for diagnosing and treating the patient, with the AI system merely providing suggestions. Ultimately, healthcare professionals make the final decision, implying that their decisions and resulting actions can still be evaluated based on established standards of medical ethics and medical law. From this perspective, there seems to be no immediate need for additional regulatory adjustments, as existing frameworks primarily regulate the actions of healthcare professionals (e.g., Gilvary et al. 2019; De Boer and Kudina 2021; Kudina and de Boer 2021; Jiang et al. 2017; Lee 2022; Morley et al. 2020; Schleidgen and Friedrich 2023; Rudschies and Schneider 2024; Speck et al. 2021).

However, as AI systems become more advanced and more integrated into the traditional relationship between healthcare professionals and patients, the consideration of AI systems within this context becomes more complex. In this context, it is quite likely that, to some extent, healthcare professionals will not be able to fully understand the AI system and, in particular, how the AI system arrives at its diagnoses and treatment proposals. This is, in some sense, even the goal and the rationale behind the use of the technology: AI systems are designed to enhance diagnostic and treatment capabilities beyond what healthcare professionals alone can achieve – if AI systems would not offer such superior analytical capacities, their use in the medical context would be questionable right from the beginning. However, there are still important reasons to assume that the capacities of AI systems are so impressive that it would be prima facie irresponsible not to use them in light of the possible positive impacts on the well-being of the patient. However, the more AI systems are capable of fulfilling their task successfully, the more they will disrupt the traditional interaction between healthcare professionals and their patients.

This disruption creates an inherent tension: If AI systems were ineffective in cancer diagnoses, it would be pointless to use them. However, the more proficient they become, the more likely it is that their capacities exceed the interpretative capacities of healthcare professionals. We have to keep in mind here that cancer treatment is not only a technical process where we have to evaluate biological processes and the probabilities of successful medical interferences. Rather, it is a process where healthcare professionals must deal with particular, individual situa-

tions of patients, their wishes, fears, hopes, and the outlook of life of their patients. Whatever AI systems suggest, we must expect that the healthcare professional is still in the position to mediate between the available medical information and the specific situation of their patients, and the interference of AI systems should support them in their tasks to arrive at diagnoses and treatments. The integration of AI systems raises at least the following two questions.

First, what do we have to presuppose about the capacities of healthcare professionals regarding the use of AI systems? To what extent are these persons capable of understanding the way AI systems work? While healthcare professionals are not expected to be experts in AI systems, they must still be capable of assessing the reliability and implications of AI-generated recommendations to evaluate the value of the retrieved diagnostic and treatment proposals. Looking ahead, as AI systems become even more sophisticated, we can ask: At what point does it become unreasonable to expect healthcare professionals to be capable of understanding why an AI system has made a particular suggestion? Of course, we cannot specify the technical details that healthcare professionals need to know. However, suppose we assume that there is a continuum between a simple layperson who is just able to use a computer and an IT expert. In that case, we should wonder at which point we would locate the healthcare professional in order to ask: Is this person still fully in charge of the diagnostic and treatment process?

Second, what do we have to presuppose about the qualities of an AI system that we would ask: This AI system is designed in a way that healthcare professionals with mediocre technical knowledge are still sufficiently capable of understanding why an AI system retrieved a specific proposal? And what kind of adjustments to AI systems are sufficient to enable the healthcare professional to have a sufficient understanding?

These questions are critical because, as long as healthcare professionals are responsible for treatment decisions, they must also remain accountable for them. Since we deal with cancer detection, it may easily be a question of life and death. This implies that healthcare professionals are fully accountable for their decisions; to emphasize the importance of this point: If something goes wrong, their decisions and actions could be assessed in the courtroom. For the understanding of healthcare professionals, we have to make realistic assumptions about the technical knowledge of a mediocre person who received training in medicine, and not AI systems. Of course, we can assume that future healthcare professionals will have more advanced technical knowledge compared to today, but it would be unrealistic to assume that healthcare professionals will be IT experts. This leads to further questions: How much expertise should be required to hold healthcare professionals accountable for their decisions? And how should an AI system be designed so that it is capable of providing the healthcare professional with the information needed to make those decisions responsibly?

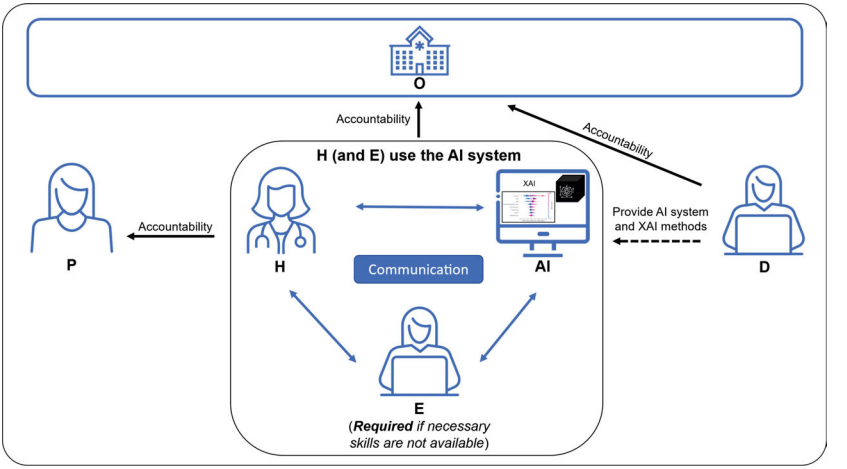
Suppose it were the case that the proposals of the AI system significantly exceed the understanding of healthcare professionals. In that case, it may be necessary to rethink the division of labor: Thus, we can think about a laboratory for cancer detection in which healthcare professionals are still in charge of their medical decisions and communication with their patients, but a clinical AI expert would be accountable for commanding the AI system and the interpretation of the suggestions of the AI system. In such cases, we come to the same questions regarding the necessary features of the AI system and the required capacities of the clinical AI experts, but we would additionally have to ask: What would be required regarding the capacities of the healthcare professionals and the clinical AI expert to communicate successfully with each other? And if they make a mistake, who will be held accountable in the courtroom? Can we determine under what conditions healthcare professionals will be in the dock, and under what conditions the clinical AI expert will be, and when both might be held accountable? If we were not capable of determining those questions on the level of general criteria, we would have a typical case of accountability diffusion, often discussed under the label 'responsibility gap' (see, e.g., Düwell 2012, pp. 185–192; Matthias 2004), thus, situations where nobody is accountable. In a situation where the life and death of persons are at stake that is unacceptable.

If we have difficulties determining those things, we could further ask whether there is a possibility to enhance the feedback procedures regarding the functioning of the AI system and whether we could install some form of general oversight of the process. Both things are necessary, anyhow, since even if accountability frameworks are well-defined at one point in time, ongoing technological advancements will necessitate continual reassessment. Thus, what are the requirements regarding the feedback procedures where the developer of the AI system (let us call them 'D') to ensure that the AI system is designed appropriately to enable healthcare professionals with mediocre technical abilities (let's call them 'H'), perhaps with the support of clinical AI experts (let us call them 'E'), to be able to evaluate the suggestions of the AI system? Additionally, what are the requirements regarding institutions that are capable of exercising oversight (let us call them 'O')? O must be in a position to judge whether the relationship between the features of the AI system and those capacities that one can reasonably expect from healthcare professionals (plus clinical AI experts) to have an appropriate accountability relationship towards the patient (let us call them 'P').

In the scheme below, we illustrate our proposed model by focusing on accountability relationships (see Figure **1**). It highlights that healthcare professionals remain ultimately accountable to their patients. In some cases, the healthcare professionals may require technical assistance from the clinical AI expert (let us call them 'E'), while in other cases, they may independently interpret the suggestions of the AI system. Developer D is only accountable for realizing the AI system in ways that are required to bring healthcare professionals H in a position to realize accountability

vis-à-vis patient P. The regulatory authority O is accountable for the oversight that the use of AI systems is regulated in a way that healthcare professionals H can fulfill their tasks vis-à-vis their patients P. In this vein, it should be emphasized that these considerations are fully independent of the content of the normative standards that are applied and the specific ethical theory that is presupposed (for an overview of possible ethical approaches, see Düwell 2012). Whatever normative standards one applies, it always has to be presupposed that healthcare professionals H are capable of applying them in a specific practice. That implies that they possess the capacity to understand the practice to the extent that they are able to interfere according to those standards.

Figure 1: Proposed model for accountability relationships



In summary, we can systematize the relevant capacities and accountability relationships:

(1) H must be capable of understanding the proposals of the AI systems, potentially with the support of E, and he or she is accountable towards P, and in case of failure of H, P must have legal possibilities to address this.
(2) E is only responsible for H. E must have the capacity to deal with the AI system competently and understand the questions of H.
(3) O must be capable of providing feedback to D regarding the appropriate functioning of AI systems. D is accountable to O. Of course, there can and should be direct communication from H and E with D, but the line of accountability goes towards O.

(4) H (and E) are accountable to O. O must be capable of assessing whether AI systems are designed in a manner that enables H to take accountability for medical decisions vis-à-vis the patient.

We have to assume that it is, in principle, possible that H, E, D, and O can end up in a courtroom because of a failure regarding their part of the accountability. Since we talk about serious decisions (decisions about life and death), it must be possible to differentiate between the respective parts of the accountability. It is particularly important to stress that the AI system must be designed, developed, and deployed in a way that the patient has the possibility to claim his or her rights in the courtroom. The only instance against which the patient can claim rights is the healthcare professional. If that were not the case, patient rights would be seriously endangered. This, however, brings the healthcare professional into a much more vulnerable position. Therefore, it is necessary to specify what healthcare professionals can be held accountable for and where this accountability ends. Without such differentiation, accountability could either be entirely absent or unfairly distributed among all parties, both of which are unacceptable outcomes. The question for the following part is, therefore, what are the possibilities from the perspective of XAI to provide us with criteria for the differentiation of the respective parts of the accountability?

4. Possible Answers to these Challenges From the XAI Perspective

In what follows, the role of XAI methods is discussed in enhancing transparency at different stages of the design, development, and use of AI systems (e.g., Molnar 2019). These methods aim to assist each actor involved in fulfilling their unique obligations by providing insights that enable explanations and justifications of decisions to their respective counterparts (e.g., the patient P or regulatory authorities O). Below, we analyze each accountability relationship presented in the previous section and discuss how XAI methods can enhance the transparency of the AI system and support the involved parties with their accountability obligations.

D is accountable to O. When developing medical AI systems, we suggest that D should not be allowed to design and develop any AI system without requirements provided by the regulatory authorities O. Thus, D should be held accountable to meet the requirements of O, as specified in how medical AI systems can be designed and developed. In particular, such requirements should include accountability obligations requiring explanation and justification about the performance of the AI system and requiring explanation and justification about the functionality of their AI systems to communicate the limits and constraints under which the AI system can be used with minimal risk of ethical issues. To support D with these accountability obligations, they need to ensure high performance (i.e., high efficiency with min-

imal error rates) of the AI system and need to understand the behavior of their AI system. This implies that D needs transparency about the training process to ensure high performance and needs transparency about the behavior of their AI systems. In this context, D can rely on XAI methods that help them to understand and debug their training data, such as influential instances, which allow D to increase the performance of their AI system. Additionally, D can rely on global model-agnostic methods, such as partial dependency plots, to comprehend the behavior of their AI system, which allows D to communicate the limits and constraints of their AI systems.

XAI methods, such as influential instance analysis (see Table **1**), allow D to evaluate the impact of specific data points on the performance of their AI system (Molnar 2019). By identifying and removing specific data points that increase the error rate of their AI system, D can better understand how their AI system learns from the given training data. This allows D to enhance their training data and increase the subsequent performance of their AI system, thereby minimizing harmful outcomes such as incorrect diagnoses and medical treatments. As a result, this process supports D in explaining and justifying the robustness of their AI system to O and fulfills their accountability obligations.

Additionally, XAI methods like partial dependence plots (Friedman 2001) and permutation feature importance (Fisher et al. 2019) provide D with a high-level understanding of the behavior of their AI system (Molnar 2019). This high-level understanding of their AI system enables D to ensure that the AI system operates as intended and to detect unintended biases. This information allows D to edit parameters or change the architecture of their AI system to ensure intended behavior. Importantly, the abstract level of explanation provided by these XAI methods does not require D to have in-depth domain knowledge of medicine, making it practical for developers to verify the behavior of their AI system and communicate its limits and constraints to O.

H (and E) are accountable to O. We suggest that O acts as the intermediary between D and H (and E), holds D accountable for overseeing the performance of the AI system, and signs the suitability of the AI system for deployment. To fulfill this accountability obligation and oversee the AI system, O needs experts in both fields (i.e., medicine and AI development) to evaluate the performance from a technical and semantic level. For the evaluation, O needs to rely on global model-agnostic methods provided by D to understand and evaluate the behavior of the AI system. These XAI methods help O check compliance with specified requirements by avoiding or mitigating potential ethical issues. Based on this assessment, O can approve or reject the AI system and provide D with additional information on how to improve their AI system.

Once O scrutinized the AI system, we suggest that H (and E) should be accountable for understanding the limits and constraints of the AI system. This ensures that

H (and E) are aware that the AI system might suggest false diagnoses or medical treatments, and that they should not blindly follow the suggestion of the AI system.

H is accountable to P. H bears the accountability for the diagnosis and medical treatment of their patients. Although H might interact with the AI system and draw on the output, H remains accountable for ensuring correct medical outcomes. Thus, H needs transparency regarding the functionality of the AI system.

By relying on global model-agnostic methods, H gains a broad understanding of the AI system and its overall behavior. Together with the help of E, global model-agnostic methods help H to assess whether the recommendations of the AI system align with their expectations and medical knowledge. However, while global model-agnostic methods help H (and E) to gain an overall sense of the AI system, H (and E) are usually confronted to explain and justify individual decisions. Thus, global model-agnostic methods are usually too broad to explain and justify the behavior of the AI system in specific situations. Consequently, D must also equip H (and E) with local model-agnostic methods, like LIME (Ribeiro et al. 2016).

Local model-agnostic methods, like LIME, offer H (and E) a detailed explanation of specific recommendations for the AI system (Molnar 2019). By analyzing individual recommendations in detail, H can determine whether the recommendations of the AI system are logical and consistent with known medical causal relationships. This fine-grained transparency empowers H to incorporate its medical expertise and make informed decisions about diagnosis and medical treatment, respectively to explain and justify its decisions when false diagnoses or medical treatments occur. This helps H to fulfill its accountability obligations, even in cases of erroneous outputs.

5. Outlook

We acknowledge that our proposed model looks complicated. There are at least four to five parties involved: patient, healthcare professional (plus perhaps a clinical AI expert), developer and oversight authorities – and, of course, the AI system in between. While the distinctions between those parties are based on functional roles, it is theoretically possible for a healthcare professional to act as their own clinical AI expert or developer, though this is unlikely in practice. We do not claim that the division of accountability does work, but we claim that the regulation of the relationships between those parties is required if it is still possible to address questions of accountability appropriately to avoid an accountability diffusion. In the case of cancer diagnoses, the disappearance of accountability implies a loss of patient rights. A patient would not be able to address his or her claim for appropriate treatment to an instance that is expected to fulfill this claim and which is able to do so.

In that context, it is important to emphasize that the model is not meant to be static. We just distinguish the different instances and their role obligations. But that

does not mean that the relationship is a static one. It is also possible that healthcare professionals learn to use AI systems as part of a practice where they are still in charge, but it is likewise possible that they become increasingly dependent on the systems and lose their own diagnostic capacities and the ability to rely on their own judgment. It is not decided in advance how the future will develop. However, the central point is that the advantages of AI systems can only be utilized at the cost of more involved parties and with more complex relationships that have to be evaluated. However, we propose that using XAI methods can help each party fulfill the accountability obligations that arise at their specific touchpoints with the AI system. In this vein, our model allows us to assess whether the advantages of the use of AI systems are worth the introduction of more complex accountability relationships and how XAI methods can help in each phase. But in any case, if it is responsible to introduce AI systems, then it must be seen as a necessary requirement that there is still a functioning system to hold human actors accountable vis-à-vis the patient to ensure patient rights. This implies that the accountability of the healthcare professional would still be the cornerstone of medical ethics and medical law. Additionally, it would still be possible for one healthcare professional to speak with one patient from person to person, listen to the wishes and fears, the hopes and expectations of this person, and on the basis of this conversation, propose to this patient the best treatment for this person. The introduction of AI systems should not undermine this conversation from one person to another. It should still be possible that the patient accuses the healthcare professional of giving wrong advice. All of this is required so that medical law and medical ethics are still functioning.

References

Adadi, A. and Berrada, M. (2018): "Peeking Inside the Black-box. A Survey on Explainable Artificial Intelligence (XAI)", in: IEEE access, 6, pp. 52138–52160.

Apley, D. W. and Zhu, J. (2020): "Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models", in: Journal of the Royal Statistical Society Series B: Statistical Methodology, 82, pp. 1059–1086.

Asatiani, A. et al. (2020): "Challenges of Explaining the Behavior of Black-box AI Systems", in: MIS Quarterly Executive, 19, pp. 259–278.

Beeravolu, A. R. et al. (2021): "Preprocessing of Breast Cancer Images to Create Datasets for Deep-CNN", in: IEEE Access, 9, pp. 33438–33463.

Berente, N. et al. (2021): "Managing Artificial Intelligence", in: MIS Quarterly, 45, pp. 1433–1450.

de Boer, B. and Kudina, O. (2021): "What is Morally at Stake When Using Algorithms to Make Medical Diagnoses? Expending the Discussion Beyond Risks and Harms", in: Theoretical Medicine and Bioethics, 42, pp. 245–-266.

Brancaccio, G. et al. (2024): "Artificial Intelligence in Skin Cancer Diagnosis. A Reality Check", in: Journal of Investigative Dermatology, 144, pp. 492–499.

Breiman, L. (2001): "Random forests", in: Machine learning, 45, pp. 5–32.

Bringsjord, S. (2008): "Ethical Robots. The Future Can Heed Us", in: AI & Society, 22, pp. 539–550.

Cooper, A. F. et al. (2022): "Accountability in an Algorithmic Society. Relationality, Responsibility, and Robustness in Machine Learning", 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea.

Cunningham, P., Cord, M. and Delany, S. J. (2008): Supervised learning. Machine Learning Techniques for Multimedia. Case Studies on Organization and Retrieval, Springer.

Düwell, M. (2012): Bioethics. Methods, Theories, Domains, Routledge.

Dwivedi, R. et al. (2023): "Explainable AI (XAI). Core Ideas, Techniques, and Solutions", in: ACM Computing Surveys, 55, pp. 1–33.

Fisher, A., Rudin, C. and Dominici, F. (2019): "All Models are Wrong, but Many are Useful. Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously" in: Journal of Machine Learning Research, 20, pp. 1–81.

Fiske, A., Henningsen, P. and Buyx, A. (2019): "Your Robot Therapist Will See You Now. Ethical Implications of Embodied Artificial Intelligence in Psychiatry, Psychology, and Psychotherapy", in: Journal of Medical Internet Research, 21, e13216.

Friedman, J. H. (2001): "Greedy Function Approximation. A Gradient Boosting Machine", in: Annals of Statistics, 29, pp. 1189–1232.

Gilvary, C. et al. (2019). "The Missing Pieces of Artificial Intelligence in Medicine", in: Trends in Pharmacological Sciences, 40, pp. 555–564.

Goldstein, A. et al. (2015): "Peeking Inside the Black Box. Visualizing Statistical Learning With Plots of Individual Conditional Expectation", in: Journal of Computational and Graphical Statistics, 24, pp. 44–65.

Gupta, M. and Gupta, B. (2018): "A Comparative Study of Breast Cancer Diagnosis Using Supervised Machine Learning Techniques", Second International Conference on Computing Methodologies and Communication (ICCMC), IEEE, pp. 997–1002.

Hariharan, S. et al. (2023): "XAI for Intrusion Detection System. Comparing Explanations Based on Global and Local Scope", in: Journal of Computer Virology and Hacking Techniques, 19, pp. 217–239.

Hunter, D. et al. (2012): "Selection of Proper Neural Network Sizes and Architectures – A Comparative Study", in: IEEE Transactions on Industrial Informatics, 8, pp. 228–240.

Jiang, F. et al. (2017): "Artificial Intelligence in Healthcare. Past, Present and Future", in: Stroke and Vascular Neurology, 2, pp. 230–243.

Krogh, A. (2008): "What are Artificial Neural Networks?", in: Nature Biotechnology, 26, pp. 195–197.

Kudina, O. and de Boer, B. (2021): "Co-designing Diagnosis. Towards a Responsible Integration of Machine Learning Decision-support Systems in Medical Diagnostics", in: Journal of Evaluation in Clinical Practice, 27, pp. 529–536.

Lee, S. S. (2022): "Philosophical Evaluation of the Conceptualisation of Trust in the NHS' Code of Conduct for Artificial Intelligence-driven Technology", in: Journal of Medical Ethics, 48, pp. 272–277.

Linardatos, P., Papastefanopoulos, V. and Kotsiantis, S. (2020): "Explainable Ai. A Review of Machine Learning Interpretability Methods", in: Entropy, 23, 18.

Matthias, A. (2004): "The Responsibility Gap. Ascribing Responsibility for the Actions of Learning Automata", in: Ethics and Information Technology, 6, pp. 175–183.

Molnar, C. (2019): Interpretable Machine Learning.A Guide for Making Black Box Models Explainable, Lulu.com, self-published.

Morley, J. et al. (2020): "The Ethics of AI in Health Care. A Mapping Review", in: Social Science & Medicine, 260, 113172.

Nassif, A. B. et al. (2022): "Breast Cancer Detection Using Artificial Intelligence Techniques. A systematic Literature Review", Artificial intelligence in medicine, 127, 102276.

Osareh, A. and Shadgar, B. "Machine Learning Techniques to Diagnose Breast Cancer", 2010 5th international symposium on health informatics and bioinformatics, 2010. IEEE, pp. 114–120.

Perincheri, S. et al. (2021). "An Independent Assessment of an Artificial Intelligence System for Prostate Cancer Detection Shows Strong Diagnostic Accuracy", in: Modern Pathology, 34, pp. 1588–1595.

Ponnada, V. T. and Srinivasu, S. N. (2019): "Edge AI System for Pneumonia and Lung Cancer Detection", in: International Journal of Innovative Technology and Exploring Engineering, 8, pp. 1908–1915.

Retzlaff, C. O. et al. (2024): "Post-hoc vs Ante-hoc Explanations. xAI Design Guidelines for Data Scientists", in: Cognitive Systems Research, 86, 101243.

Ribeiro, M. T., Singh, S. and Guestrin, C. (2016): "'Why Should I Trust You?' Explaining the Predictions of Any Classifier", Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144.

Rudin, C. (2019): "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead", in: Nature Machine Intelligence, 1, pp. 206–215.

Rudschies, C. and Schneider, I. (2024): "Ethical, Legal, and Social Implications (ELSI) of Virtual Agents and Virtual Reality in Healthcare", in: Social Science & Medicine, 340, 116483.

Russell, S. J. and Norvig, P. (2016): Artificial intelligence: a modern approach, Pearson.

Schleidgen, S. and Friedrich, O. (2023): "Künstliche Intelligenz in Medizin und Pflege", in: Ethik in der Medizin, 35, pp. 169–172.

Sechopoulos, I., Teuwen, J. and Mann, R. (2021)"Artificial Intelligence for Breast Cancer Detection in Mammography and Digital Breast Tomosynthesis. State of the Art", in: Seminars in Cancer Biology, pp. 214–225.

Shin, D. (2021): "The Effects of Explainability and Causability on Perception, Trust, and Acceptance. Implications for Explainable AI", in: International Journal of Human-Computer Studies, 146, 102551.

Shravya, C., Pravalika, K. and Subhani, S. (2019): "Prediction of Breast Cancer Using Supervised Machine Learning Techniques", in: International Journal of Innovative Technology and Exploring Engineering (IJITEE), 8, pp. 1106–1110.

Speck, H. et al. (2021): Digitalisierung im Gesundheitswesen: anthropologische und ethische Herausforderungen der Mensch-Maschine-Interaktion, Herder.

Suzuki, H. et al. (2021): "Artificial Intelligence for Cancer Detection of the Upper Gastrointestinal Tract", in: Digestive Endoscopy, 33, pp. 254–262.

Tosun, A. B. et al. (2020): "Explainable AI (xAI) for Anatomic Pathology", in: Advances in Anatomic Pathology, 27, pp. 241–250.

Zheng, D., He, X. and Jing, J. (2023): "Overview of Artificial Intelligence in Breast Cancer Medical Imaging", in: Journal of Clinical Medicine, 12**,** 419.