Ethics and Information Technology (2024) 26:26
https://doi.org/10.1007/s10676-024-09764-8

ORIGINAL PAPER

Percentages and reasons: AI explainability and ultimate human responsibility within the medical field

Markus Herrmann[1] · Andreas Wabro[2] · Eva Winkler[2]

Published online: 9 April 2024

Abstract
With regard to current debates on the ethical implementation of AI, especially two demands are linked: the call for explainability and for ultimate human responsibility. In the medical field, both are condensed into the role of one person: It is the physician to whom AI output should be explainable and who should thus bear ultimate responsibility for diagnostic or treatment decisions that are based on such AI output. In this article, we argue that a black box AI indeed creates a rationally irresolvable epistemic situation for the physician involved. Specifically, strange errors that are occasionally made by AI sometimes detach its output from human reasoning. Within this article it is further argued that such an epistemic situation is problematic in the context of ultimate human responsibility. Since said strange errors limit the promises of explainability and the concept of explainability frequently appears irrelevant or insignificant when applied to a diverse set of medical applications, we deem it worthwhile to reconsider the call for ultimate human responsibility.

Keywords Explainability · Interpretability · Responsibility · Black box · AI

Introduction

Currently, the trend goes towards understanding AI as a non-autonomous instrument embedded in a context of human agency. For instance, in the medical field AI is celebrated as the 'stethoscope of the 21st century' (Subodh, 2023; Mesko, 2017), i.e., as an instrument that is as mundane as it can be.

Within the ethical discussion, two demands mirror such an understanding: first, that in high-risk settings human oversight of an AI tool should be paramount, and second, that the way an AI generates its output should be "explainable to those directly and indirectly affected" (High-Level Expert Group on Artificial Intelligence, 2019, pp. 13 and 16).

Here, the discussion of medical ethics is of a special character. Not only are the stakes extremely high, but these two aspects of human agency are radically condensed into one person: the physician. It is she who should be ultimately responsible for a course of action that is assisted by an AI system, and it is she to whom an AI output should remain explainable (and only via her to the patient).

The issues of ultimate human responsibility and explainability were addressed by several influential institutions in their public statements[1] and they have led to an intense discussion within the medical research community: Notably, London (2019) stirred up a controversy as he argued against the explainability constraint in favor of higher accuracy (with Babic et al., 2021 and Da Silva, 2023 following him). Among his critics, Grote and Berens (2020) have stressed that the epistemic situation a black box AI creates for the

✉ Markus Herrmann
markus.herrmann@nct-heidelberg.de

[1] National Center for Tumor Diseases (NCT), NCT Heidelberg, a partnership between DKFZ and Heidelberg University Hospital, German Cancer Research Center (DKFZ) Heidelberg, Division Applied Tumor Immunity, Heidelberg University, Medical Faculty Heidelberg, Heidelberg University Hospital, Department of Medical Oncology, Section Translational Medical Ethics, Heidelberg, Germany

[2] National Center for Tumor Diseases (NCT), NCT Heidelberg, a partnership between DKFZ and Heidelberg University Hospital, Heidelberg University, Medical Faculty Heidelberg, Heidelberg University Hospital, Department of Medical Oncology, Section Translational Medical Ethics, Heidelberg, Germany

[1] For example, cf. High-Level Expert Group on Artificial Intelligence (2019), pp. 12–13, and 18; ZEKO (2021), pp. A11-A12; Deutscher Ethikrat (2023), pp. 26, 44, 53, 142–144 and 161–162; Ontario (2022).

physician is at odds with the idea of her ultimate human responsibility.

This article provides an in-depth analysis of the epistemic situation of a physician that is using an AI system in her decision-making process, especially a black box AI. As we will see, however, the situation is even worse than the proponents of an explainable AI (a so-called XAI) think: Strange errors and non-localized data processing of AI solutions decontextualize AI output from human understanding. Yet, the variety of medical AI applications as well as the fact that even an XAI cannot fully solve the dire epistemic situation that is created by the threat of strange errors suggest that the question of a physician's ultimate responsibility cannot be answered by a simple call for explainability. In contrast, it is argued for the thesis that systematic normative and translational clinical research is necessary to develop case-sensitive principles and procedures concerning the scope and limits of ultimate human responsibility in the context of medical AI implementation.

It is important to highlight that this is not tantamount to rejecting the call for explainable AI systems within the medical field, as there are other arguments speaking in favor of it apart from the epistemic situation that it creates for the physician (for an overview, cf. Da Sliva, 2023, p. 6). Although this article provides an in-depth examination of the epistemic situation created by an AI lacking explainability, it is important to highlight that its main objective is *to evaluate the call for ultimate human responsibility* and the role of the physician as the end user of an AI system.

Key concepts

In this article, 'ultimate human responsibility' refers to a physician being responsible for a diagnosis or treatment in which an AI system was employed. This shall be understood within the scope and notions of *moral* responsibility.[2] The concept of ultimate human responsibility shall not be reduced to its jurisprudential aspects of *legal liability*, which have already been and continue to be highly discussed among stakeholders.[3]

Here, the term 'ultimate' needs further clarification. What is the difference between 'ultimate human responsibility' and mere 'human responsibility'? Although the term 'ultimate' is used in public statements, it is not explicitly defined. However, it is commonly used in the context of human oversight, i.e., that there is no automated decision making, but that humans are always on or in the loop (e.g., UNESCO, 2024 and ZEKO, 2021, pp. A11-12). Hence, mere 'human responsibility' refers within this article to the idea that AI cannot bear responsibility and that the responsibility for its use should be located with humans (for instance, the responsibility a manufacturer bears for the robustness of her devices). In contrast, 'ultimate human responsibility' will be much narrower understood: it expresses the idea that the *human end user* is responsible for countering errors that might occur – and that is in the medical field: the physician.

Further, it is necessary to distinguish the process of *ultimate decision making* or *human oversigh*t on the one hand, and the fact of bearing *ultimate responsibility* for this decision on the other hand, for only the latter necessarily implies the normative concepts of praise- and blameworthiness. For instance, by means of human oversight, reaching a "final decision" in a complex medical case of diagnosing cancer does not per se call for the ascription of praise or blame to that person's decision. Although the process of decision making frequently results in human conduct to be further assessed and evaluated, deciding on a particular medical issue does not necessarily require the decision maker to carry responsibility for this decision in every aspect, even more so when considering an AI algorithm's influence on this human's decision-making process.

The second key concept of this article is the concept of explainability. According to proponents of explainability, it is not sufficient that an AI calculates accurate output. They argue that humans should be able to understand *why* the respective output was generated. Especially in its clinical application, humans should be able to comprehend the specific causes that led to an AI generated diagnosis or treatment recommendation. It is, e.g., not enough that an AI indicates the presence of a tumor, it also needs to present *why* it indicates it.

This explainability constraint is in direct opposition to the black box nature on which many AI tools operate (Zednik, 2021, p. 265), for some are programmed in a way that even their creators cannot reproduce how the system exactly obtained a specific result.

It is important to notice that the term explainability is inconsistently used within the discussion about AI systems (Zednik, 2021, pp. 268–269; Ursin et al., 2023, pp. 179 and 184). In addition, similar terms such as 'interpretability' and 'explicapability' are commonly, but not necessarily used. For the purpose of this article, a system is considered to be explainable if its ordinary end user (i.e., in this context, the physician) can understand how the algorithm arrived at its result in a given case.

Finally, this article draws heavily on a conceptual distinction from philosophical epistemology, namely the

[2] Importantly, it shall not be confused with previously discussed aspects of 'ultimate responsibility' in deterministic or compatibilist disputes of philosophical ethics (see e.g., Strawson, 1994).

[3] For a detailed analysis of different legal liability schemes in an AI context, such as vicarious or strict liability (cf. e.g., Wendehorst, 2022).

distinction between *first-order evidence* and *higher-order evidence* (Kelly, 2010).

To illustrate those terms: First-order evidence for the fact that it has rained could be, for example, that I perceive that it is raining. Higher-order evidence, on the other hand, would be my friend telling me that it has rained. The latter evidence could not exist without former first-order evidence. Therefore, higher-order evidence is evidence of further evidence. It often presents itself as someone making an assertion, where we assume that this person has some further evidence as justification for her claim.

To further elaborate on this distinction, we want to add an additional concept: In a reflective process, higher-order evidence can be addressed like first-order evidence by refuting or affirming reasons. For instance, the first-order evidence that the street is wet could be refuted by the argument that the fire department held a drill there. Similarly, the higher-order evidence that my friend told me could be refuted by the claim that my friend is a notorious liar when it comes to talks about weather. We will call such reasoning *higher-order evidence addressing reasons* (or short: HEAR).

An asymmetry of first- and higher-order evidence

Grote and Berens (2020) have argued that the output of a *black box AI* cannot rationally be resolved with a conflicting diagnosis of a physician. They highlight that in case of conflicting diagnoses of a physician and a black box AI, the AI output is similar to such higher-order evidence. For example, if an AI indicates the presence of a melanoma, it is similar to a physician telling a colleague that the suspected lesion at hand is a melanoma. However, a crucial difference is that if a physician disagrees with a colleague, she can start arguing with him. She can present her first-order evidence (like presenting the pattern of the lesion) that led to her conclusion. And she can ask him for his first-order evidence, where he might present similar patterns that are known to be non-malignant.

However, a black box AI for melanoma detection does not offer such reasoning (i.e., first-order evidence), but simply indicates the presence (or absence) of a melanoma. This could be compared to a physician who makes a diagnosis, but refuses to provide any reasons for it if asked by colleagues or patients. In such a situation, the disagreement among peers cannot be rationally resolved.

The only way of evaluating the output of a black box AI is to consider its performance in previous cases. Most of the time, that means relating to its *overall success* rate only (in part 4, we discuss other ways as well). However, such overall success rates comprise percentage values and merely *corroborate* the higher-order evidence (namely, the validity of the black box AI output). In this regard, they are only higher-order evidence addressing reasons (HEAR). Importantly, they do not resemble first-order evidence and cannot be argued with by presenting first-order evidence.

According to Grote and Berens (2020, p. 208), the incompatibility of evidence created when using a black box AI leads to the ethical problem that a physician could be accountable for the final diagnosis but would not have the epistemic means to evaluate it.

First-order evidence and success rates

In the following, it is argued that the situation of a physician using a black box AI is even worse than Grote's and Berens' analysis indicates. For this, we need to start with an in-depth analysis of all the evidence that is available to a physician when in conflict with a black box AI decision support system.

Speaking in favor of her diagnosis, a physician has several types of first-order evidence: For instance, she might have access to radiological or histological images for proper examination (these are the images that are also presented to the AI). But her first-order evidences actually go far beyond that: context knowledge, testing results, and other information she retrieved by taking a patient's history and examining the patient herself.

On the other hand, there is higher-order evidence in form of an AI output that is speaking against her diagnosis and that is backed-up by its overall success rate.

Still, there is also higher-order evidence speaking in favor of the physician's initial judgement. This is similar to the overall success rate of the black box AI: the trust in her capabilities as an experienced physician that has evolved over time and results in knowledge about the limitations of her capabilities.

For the sake of the argument, we assume the favorable case that the physician even knows her own overall success rate. One might now think that this is sufficient information to resolve the issue: The higher success rate should take precedence over her trust in her professional capabilities when it comes to the final diagnosis. However, the situation is complicated by the fact that an AI makes *different* mistakes than physicians (ZEKO, 2021, p. A3). It is not the case that the person (or machine) with the lower success rate makes the same mistakes as the other side, plus some additional ones. On the contrary, a higher success rate does not preclude that mistakes occur that the other side would not have made.

Because of this disparity in mistakes, we need to distinguish between four cases.

Compared to a physician, a black box AI could generally be:

a) less often correct.
b) equally correct.
c) more often correct.
d) always correct.

Of these four cases, case *d* has an obvious rational resolution: If the AI is always successful, it is reasonable for the physician to follow its diagnosis all the time. If the AI is less successful (case *a*), it seems that there is a rational resolution possible as well. The higher-order evidence available to the physician speaks in favor of her diagnosis, for she has the better overall success rate. And as there is no first-order evidence that contradicts her (because a black box AI does not provide her with such evidence), the only first-order evidence available to her speaks in her favor as well.

The case of an equally successful black box AI mirrors the cases of the original philosophical discussion of peer disagreement which has led to Thomas Kelly's distinction between first- and higher-order evidence (Kelly, 2010). Here, the higher-order evidences seem to cancel each other out. As Kelly has argued, in such a situation it is still rational for a person to stay with her first-order evidence, although such an impasse of higher-order evidence might be sufficient reason to at least reexamine her first-order evidence. If she then cannot find anything that deserves more attention, she is justified to go on with her initial judgement.

This leaves us with case *c*: An overall more successful black box AI contradicts a physician's judgement. Here, it is important to stress that this is the *default scenario.* This is the most relevant case in the discussion of clinical AI implementation, for impressive success rates of AI applications (and the prospect of their further increase) are one of the main driving forces of ongoing debates (e.g., London, 2019, p. 18; Grote & Berens, 2020, p. 205; Da Silva, 2023, p. 1).

How can we resolve the conflict in this case? The higher-order evidence of the physician and of the algorithm do not seem to cancel each other out but seem to speak, to a certain degree, in favor of the AI diagnosis. Yet, on the other hand there is the first-order explanation that led to the physician's initial judgement. As we have seen above, there is no rational resolution between first-order and higher-order evidence. Apparently, only the physician's own higher-order evidence and the one of the black box AI can be directly rationally resolved into action. Of course, the physician could ask a colleague for a second opinion, but this does not provide an answer to the conundrum of how to rationally resolve the first-order evidences with the black box AI output. It seems that only higher-order evidence could have such a resolution.

However, as we see in the next part, there is an important epistemic difference between a physician and an algorithm when it comes to higher-order evidence, rendering even such a resolution impossible.

Context and higher-order evidence addressing reasons

At a first glance, the success rate of a physician and that of a black box AI seem comparable. If an AI used in a setting of diagnostic decision support is correct in 95% of cases, and a physician is correct in 85% of such cases, this appears to be information of the same category. But a real-world situation is a lot more complicated. For, there is not only an asymmetry in the availability of first-order evidence, but in the nature of the higher-order evidence. Precisely, there is an asymmetry in the availability of higher-order evidence addressing reasons (HEAR).

At first glance, such reasons can appear for each side. As mentioned above, a physician might know that a black box AI is biased against the patient's social strata. In such a case its overall success rate of 95% is put into a certain perspective. It is no longer possible to compare the physician's success rate of 85% to the one of 95%.

On the other hand, a physician automatically possesses a plethora of HEAR pertaining to her own success rate. Among others, these comprise knowledge about her current condition (i.e., how alert and attentive does she feel, how much clarity of mind does she have?), knowledge about the data that is accessible to her (i.e., are the images substantively differing from those she has seen before and she is comfortable with, maybe because a new imaging device is in use?), and knowledge about the difficulty of the case (i.e., does the case at hand comprise a standard case – is it as close to a textbook example as possible or is it a more difficult one?).

Specifically, strong HEAR could arise from knowledge about the data. A physician can have evidence that stems from her context knowledge and her immediate examination of the patient. This evidence is unavailable to the black box AI. The AI might have evidence of the patient's medical history and data about the previous patient contacts as well as any tests that were run on the patient. However, the physician alone has had the opportunity to gather more detailed information from examining her patient: body odor, skin and scleral color changes, sweat and agitation, movement anomalies, sensual information of a performed palpation etc.

In a given case where the physician's diagnosis depends on such information and she knows that this information is unavailable to the AI, an especially strong kind of HEAR is established. As such HEAR might render the AI's overall success rate irrelevant, it could completely invalidate the AI output. We will further call such strong HEAR *invalidating HEAR*.

Lastly, in case that an image is corrupted by breathing distortions, or there are problems with the contrast medium, there is a special kind of HEAR addressing the available data. This type might be relevant to either side's overall success rate. In such cases, the physician as well as the AI operate outside of their comfort zone.

If we compare the availability of HEAR for the physician and for a black box AI, we can find an asymmetry that makes the epistemic situation for the physician even more precarious than initially thought. We have already discussed three sources for HEAR: knowledge about the physician's current condition, knowledge about the data of the individual case, and knowledge about the difficulty of the case.

The first kind of knowledge is the least problematic: Usually, a physician can judge her own condition and an AI is per se incapable of having or experiencing a 'bad day'. But as soon as we investigate knowledge about data there is a first asymmetry: A physician knows what kind of data she is familiar with. But regardless of her education, experience, and training, she has not had the chance to acquire a comparable amount of knowledge that is entrenched in the training data of the AI. Particularly, she has not seen the vast data sets the black box AI has been trained on, and therefore she cannot know when an AI eventually runs into a distribution shift, i.e., when a case is relevantly different to the training data (e.g., because the lighting conditions are different). This also pertains to the last category: knowledge about the fact how difficult a case is to judge. Normally, a physician should be able to judge how close a case is to a textbook example. However, the situation is different for a black box AI. As mentioned above, an AI makes different mistakes than a human being does. Tasks that are easy for a physician might be extremely difficult to achieve for an AI, as it sometimes makes bafflingly strange errors. Due to the existence of such strange errors, a modern AI works outside a beneficial context that is meaningful to the physician. It hence creates an asymmetry of HEAR that casts strong doubts on a straight comparison of the overall success rates.

To understand this asymmetry better, we need to go into more detail about the nature of strange AI errors and the epistemic situation they create.

Strange errors and inaccessible higher-order evidence addressing reasons

Sometimes, an AI makes strange errors and states that its result is accurate with a very high probability, e.g., when it identifies a dragonfly as a manhole cover and states a 99% probability for this calculation (Rathkopf & Heinrichs, 2023, p. 7). These strange errors do not align what humans would consider a high difficulty of the respective case, i.e., that the dragonfly can only be poorly recognized by human means. From situations that appear simple to human observers, but still provoke the AI to make mistakes, it can be deduced that an AI apparently has to face different challenges than a human. This is not in itself a problem. If a physician knew that the case at hand is an easy textbook case to her, but hellishly difficult to the AI, it would be all the better: Such HEAR would mean that it is very likely that in this current case, the physician herself performs above her overall success rate, whereas the AI performs significantly below its rate. Therefore, it would be *easier* for her to come to a final decision.

The problem is that the strangeness of these mistakes only partially consists in their radical quality. In addition, it *takes us by surprise* to categorize a dragonfly as a manhole cover. We could not have predicted such a bizarre misclassification.

If it is true that such strange AI errors cannot be predicted, our physician is in an even epistemically more challenging situation than the asymmetry of first- and higher-order evidence initially indicated. She lacks HEAR about the difficulty of a case for a black box AI, whereas she most likely has a good understanding of how difficult it is for her. Her own overall success rate is put in context by her HEAR, but the one of the AI is not.

But is it true that humans cannot predict black box AI misclassifications? There are in fact studies where humans were able to predict such misclassifications (Zhou & Firestone, 2019; Nartker et al., 2023).

However, it is questionable whether these findings can be transferred to the medical context. Although Nartker at al. (2023) even claim that their study was conducted with AI applications bearing radiology in mind, its design deviates from such a setting significantly.

First, the sample of images for classification in these studies had quite a high variety in image modality. For example, some images were close-ups of their objects, others portrayed their objects from a normal distance. Some images were monochrome, others were polychrome. In addition, the objects on the images were situated in different semantic contexts (like a bus in a snowstorm in contrast to a bus in a bus parking lot). It might be less surprising that an AI misclassifies an image of bubbles when the image is in fact a monochrome close-up image of bubbles in coffee foam (Nartker et al., 2023, p. 2).

There are also differences in modality when it comes to radiological images. For instance, there is a variety of radiological devices potentially changing image data. But these differences are significantly harder to detect from human perspectives when compared to those mentioned above.

A second difference to this study is the fact that there are medical settings where AI output is only binary. There, an AI is not supposed to provide information on what an image portrays (a liver, a kidney etc.), but, e.g., to solely indicate whether a tumor is present or not. Here, the epistemic difference is stark: In the case of a salt shaker, there is a much wider space for error due to the lack of a preset category of classification.

Lastly, and this is by far the most important epistemic difference between the studies conducted and the medical setting, the latter is characterized by *uncertainty*. A physician is in a situation where she should counter an AI mistake while at the same time it is not obvious to her what the correct diagnosis or treatment option is. One of the reasons for implementing medical AI is that a physician's judgement is significantly prone to error. In contrast, the study participants knew what they saw when an image of a bus in a snowstorm was presented to them. The AI was not there to *meaningfully assist* them in classifying the objects on the images. Instead, from the starting point of relative certainty, they could decide on whether an AI misclassified an image or not.

All these reasons make it much more difficult in the medical field to recognize an AI error. The situation gets even worse when we turn to specific types of AI error: adversarial examples. Adversarial examples are created by intentionally modifying the input of an AI to generate a misclassification (Freiesleben, 2022, p. 8). For instance, there is the case where white noise was induced into the image of a panda, which led the AI to misclassify it as an image of a gibbon (Buckner, 2020, p. 731).

It is questioned whether such adversarial examples do even belong to the same category of misclassifications as the one where a dragonfly is misclassified as a manhole cover (Nartker et al., 2023, p. 3, Freiesleben & Grote, 2023). The most important difference is that they are *alterations* of existing images (Nartker et al., 2023, p. 3), e.g. by inducing white noise. As such, one can hope that they are not found "in the wild" and that physicians do hence not encounter misclassifications that are similar to adversarial examples.

If it happened that a physician encountered something like an adversarial example, it would be close to impossible for her to find any cause for the misclassification. For, the causes of the misclassification can be imperceptible to the human eye (Buckner, 2020, p. 731, Freiesleben, 2022). Some even consider imperceptibility to be definitional for adversarial examples (Verma et al., 2020). If there are imperceptible causes for AI errors, these causes are most unlikely to become HEAR.

It is important to highlight that there is intense research on detection tools for adversarial examples (and also for so called domain shifts that we will shortly address). However, they are far from perfect, and they might even reduce the performance rate of an algorithm (Freiesleben & Grote, 2023). Once more, this means that it would ultimately be up to the physician to decide whether there is an AI error.

Now, can such adversarial examples occur in a clinical setting? The fact that they are *intended* alterations should not be overestimated: Most are found coincidentally anyway (Freiesleben, 2022, pp. 88–91 and 96). According to Freiesleben, adversarial examples are minor domain shifts, i.e., they are cases where the input data differs relevantly from the training data, but not so much that the difference is visible to the human eye (Freiesleben, 2022, p. 93). Because the data is relevantly different to the training data, the AI misclassifies the input data. Even if such domain shifts are not the cause for adversarial examples, in the medical field there are plenty of sources for such misclassifications caused by domain shifts. For example, images can be taken with a different camera, imaging device, or applying a different contrast medium (cf. Finlayson et al., 2021).

In contrast, humans usually maintain good classification skills when classifying images taken with different cameras. Therefore, a human end user might be ignorant of those inherent classification difficulties that an AI might face. But even if a physician knew that this is a source of AI misclassification, it might still be difficult or even impossible for her to spot the difference between imaging devices. Most importantly, she does not *know* the entire training data set and therefore cannot even detect such differences, even if they are perceptible.

In sum, the reduced modality of medical images, the often binary nature of diagnostic settings, and especially the uncertainty of the physician's position as well as the problem of imperceptible (or near imperceptible) causes of AI misclassification are valid reasons that put the physician in a significantly different position when judging the difficulty of a case for an AI compared to a physician.

Black box and ultimate human responsibility

The previous considerations show that a physician consulting a black box AI can find herself in a rationally irresolvable situation if the AI output contradicts her diagnosis. Two epistemic asymmetries characterize the use of a black box AI as a decision support system in the medical field:

1. First, there is the asymmetry of first-order evidence: The physician does not have an explanation for the AI output but does know the reasons for her diagnosis.
2. Second, there is an asymmetry when it comes to higher-order evidence addressing reasons (HEAR). Although the physician has higher-order evidence that his judgement as well as the contradicting AI output is correct (namely, the overall success rate), the physician is in a less fortunate position when contextualizing the overall success rate of the AI.

Considered in isolation, these two asymmetries are not necessarily an ethical problem, but merely epistemic in nature. Yet, as soon as we presuppose the context of human agency, that is perceiving an AI as the 'stethoscope of the 2^{1s}t century', a black box AI does indeed pose an ethical problem. How can we then call for ultimate human responsibility, when we at the same time deprive a human operator from the epistemic means to live up to this responsibility?

There are two ways out of this unacceptable situation: First, one could conceive AI as a novel technology that should not be domesticated in the context of individual human agency. In this case, responsibility should be rather located with the manufacturer and the institutions of the health care system.

Second, one could pry open the black box and demand explainable AI output. XAI could provide the benefit of delivering higher-order evidence as a justification for a specific course of action, and could additionally offer comprehensible first-order evidence a physician could then equilibrate with her own reasoning.

Which of these two ways should be taken? The route of explainability seems to be the more appealing option. Not only is it in concordance with the public guidelines of several influential institutions (cf. footnote 2), but might lead to higher patient acceptance, since a physician can then present patients with further explanations for her diagnosis. In addition, it is the least radical change to a well-established practice. Ultimate responsibility remains with the physician, and AI takes the place of another medical instrument embedded in the vast array of medical tools currently deployed in practice.

In the last part, we want to question such a straightforward approach and advance the position that there is a demand for systematic research of the issue. If we examine individual cases and the different stakeholders of AI implementation, we will see that they require different approaches. Thus, we do not need a simple rejection or confirmation of ultimate human responsibility, or XAI. Instead, it is necessary to systematically develop principles that distinguish between different types of AI implementation.

Different scopes of responsibility

This part is divided into two sections. First, it casts doubt on whether explainability tools can actually enable a physician to take ultimate responsibility for a diagnosis involving an AI. Our main focus will rest on the paradigmatic case of AI use in medicine: image analysis in a pathological or radiological setting. Here, it will be shown that the strange errors which create an incompatibility of the higher-order evidence between humans and machines also push explainability tools to their limits. Such tools do not allow to reliably distinguish between strange errors and previously unknown correlations.

Second, medical AI applications beyond image analysis need to be considered. Here, it will become apparent that explainability is not necessarily desirable – or that it is at least an implausibly strong requirement.

When talking about explainable AI in the medical context, what easily comes to mind is a standard decision support system case. For example, a radiological or histological image is taken, and biomarkers are collected. An AI is supplied with this data and indicates the probability of a certain disease to be present. A radiologist or pathologist will validate this output and pass it on to a physician, who discusses it with the patient.

Even in such a paradigmatic case, ultimate human responsibility should not be presumed uncontested, since it is doubtful whether first-order evidence delivered by an explainable AI would be sufficiently comprehensible. Let us assume the case where the explainability of an AI system is achieved by using a technique analogous to heat maps (Zednik, 2021; Ghassemi et al., 2021). For example, the AI does not only indicate whether a section of tissue is a tumor, but also provides an image where parts of the tissue are colored depending on how relevant they were for the verdict of the AI. In a nutshell, the AI provides additional information by specifying *which* data led to its output.

Let us now assume that a physician receives such a heat map where even parts of the tissue far off the tissue suspected of resembling a tumor are marked as evidence for a tumor. The physician could now easily dismiss such evidence as a strange error, like when an AI mistakes a dragonfly for a manhole cover. How could a far-off region of the target tissue be an indicator? The problem is that the AI could potentially indicate a *previously unknown correlation*.

As Buckner (2020, p. 734) has pointed out, an AI's strength partially consists in its *non-localized* data processing capabilities. What does this mean? We can understand this if we turn to the problem of shortcut learning. Compared to a human, an AI might consider more parts of an image being of special relevance in an image recognition setting. For instance, there is the prominent case where an

AI classifies a patient as having melanoma because of the physician's markings on the patient's skin (Winkler et al., 2019). Here, the AI's focus was much wider than the physician's one, and the latter would have most likely focused on the localized lesion only.

However, such non-localized data processing might show benefits as well. In fact, it has the potential to discover correlations that a human observer is totally ignorant of. For instance, as severe chronic diseases cause system-wide complications, this offers the real chance to discover previously unknown correlations for said diseases.

This has significant implications for our example above. When an AI indicates that far-off parts of the location of concern might be causal for its classification recommendation, how is a physician then supposed to know (and take responsibility for) whether the AI now marks a previously unknown correlation or a bizarre error? As has been mentioned by Zednik (2021) and Ghassemi et al. (2021), heat map explainability features only highlight *what* data has been considered, not *why*. So far, there is no *causal* explanation provided of how a certain disease leads to the specific phenomena highlighted on the heat map. Neither is there anything analogous to a differential diagnosis, which is a line of reasoning that eliminates competing hypotheses by an inference to the best explanation for a specific set of symptoms (Ursin et al., 2023).

But what if the AI's explainability features were expanded and a physician could additionally obtain comparative cases from the training data set of the machine (like in Jacobs et al., 2021)? Would it not now be possible for the physician to recognize relevant changes of a far-off region in case of a tumor? Would this not allow him to distinguish previously undetected correlations from strange errors? Here, even this could not yield causal knowledge. If a significant number of cases in the training data showed something like what can be seen in the area highlighted on the heat map, we might ask whether this is due to mere chance or a significant correlation? Investigating this might be a highly important *research* task. In fact, revolutionizing research is one of the most pertinent reasons for explainable AI (Zednik & Boelsen, 2022). But as such, it can hardly be the attending physician's responsibility.

However, this does not even take into account that there might be imperceptible reasons for an algorithm's classifications (see part 5). In such a case, the heat map would highlight parts where there is nothing to detect for the human eye. Here, a comparison with images from the training data could by no means provide a definite answer as to the correct classification. Again, due to the lack of sufficient HEAR about what constitutes a difficult case for the AI, the physician is left with only one means to decide whether the AI's output is a strange error or a previously unknown correlation: the AI's performance in previous cases. In other words, she faces the same rationally irresolvable solution as before.[4]

Thus, even with an explainable AI the situation is far from similar to the case where a physician can discuss diverging diagnoses with a colleague. A colleague could make it explicit, if he thought that there is a correlation that our physician does not know. The conflict could be rationally resolved more easily. However, this is not the case with AI, even an XAI that draws on the heat map technology (and goes beyond it by presenting comparative cases).

From here, two pathways present themselves: First, one could demand that AI is designed in such a way that its processing procedures are not only explainable, but sufficiently predictable by humans, *especially by the physician*. This raises the question whether such systems can achieve the same accuracy as the systems discussed in this article (cf. London, 2019). This is a question that most likely needs to be addressed for numerous implementation settings individually.

When we now turn to cases of AI implementation that are beyond the paradigmatic ones we have discussed up to the current point (e.g., image analysis in a wider diagnostic context), a second alternative becomes more and more attractive: limiting the ultimate human responsibility of the physician.

In a diagnostic setting, image analysis is characterized by nearly exclusive cognitive activity, a comparatively open time frame (even when considering the constraints of a strained health care system), and the potential for intense communication, even for a second opinion. However, there are settings without such luxurious circumstances. Consider the case of AI based decision support within the context of open-heart surgery. For example, there currently is an AI developed recommending stitching patterns during mitral valve repair. The current standard of care is to use a template valve to determine the size of the valve and the stitching pattern (Sharan et al., 2020). In such a case, the activity is not exclusively cognitive, but involves a lot of *craftsmanship* to a relevant extent. Choosing the right stitching pattern cannot become the topic of a lengthy discussion with a colleague during the surgery, for it mainly consists in non-propositional capabilities and is embedded in a very time-sensitive environment of an open-heart surgery performed in minutes. In addition, there is no need to explaining the AI output to a patient intraoperatively. It seems odd to demand the same level of explainability (and responsibility) in this case as compared to the image analysis case mentioned above, for first-order evidence provided by the AI at hand could only be utilized to a very limited extent.

[4] This presupposes that it is not possible to invalidate the AI output by gathering more evidence that constitute invalidating HEAR.

Such a demand is even more implausible if the AI output provides novel information for decision making rather than replaces existing diagnostical methods. For example, there is an AI tool in development for detecting blood perfusion of an anastomosis during surgical procedure (Wirkert et al., 2016; Nickel et al., 2023). Currently, a surgeon has no other means to detect such a perfusion than to check his stitches and then hope for the best (Wirkert et al., 2016, p. 909 f.). This means that the AI solution provides first-order evidence currently not otherwise obtainable (how the AI processes this evidence remains non-explainable though). It could be argued that a demand for the explainability in such a case (and for the responsibility of a physician to counter AI errors) should be by magnitudes lower than in the image analysis case mentioned above, for there is no relevant alternative to it.

All these cases show that the question of AI explainability and ultimate human responsibility is not one of black and white, but one that does know gradients. Normative principles and procedures need to be developed which define the standard of care: How to use AI, how to respond in cases of deviating diagnoses (e.g., considering whether there is invalidating HEAR), and when to acknowledge that the matter is beyond the physician's reach, and she has done her duty in due diligence? Thus, research in translational clinical ethics is required to determine case sensitive principles of ethical AI use in the medical professions.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declaration

Competing interests The authors have no relevant financial or non-financial interests to disclose.

References

Babic, B., Gerke, S., Evgeniou, T., & Cohen, I. G. (2021). Beware explanations from AI in health care. *Science*, *373*(6552), 284–286. https://doi.org/10.1126/science.abg1834.

Buckner, C. (2020). Understanding adversarial examples requires a theory of artefacts for deep learning. *Nature Machine Intelligence*, *2*, 731–736. https://doi.org/10.1038/s42256-020-00266-y.

Da Silva, M., & Explainability (2023). Public Reason, and Medical Artificial Intelligence. *Ethical Theory and Moral Practice*. May: 1–20.

Deutscher Ethikrat (2023). *Mensch und Maschine: Herausforderungen durch Künstliche Intelligenz*. Berlin.

Finlayson, S. G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I. S., & Saria, S. (2021). The clinician and dataset shift in Artificial Intelligence. *New England Journal of Medicine*, *385*(3), 283–286. https://doi.org/10.1056/NEJMc2104626.

Freiesleben, T. (2022). The Intriguing Relation between counterfactual explanations and adversarial examples. *Minds & Machines*, *32*, 77–109. https://doi.org/10.1007/s11023-021-09580-9.

Freiesleben, T., & Grote, T. (2023). Beyond generalization: A theory of robustness in machine learning. *Synthese*, (109), 202. https://doi.org/10.1007/s11229-023-04334-9.

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health*, *3*(11), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9.

Grote, T., & Berens, P. (2020). On the ethics of algorithmic decision-making in healthcare. *Journal of Medical Ethics*, *46*, 205–211. https://doi.org/10.1136/medethics-2019-105586.

High-Level Expert Group on Artificial Intelligence. *Ethics Guidelines for Trustworthy AI* (2019). European Commission https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html.

Jacobs, M., Pradier, M. F., McCoy, T. H. Jr, Perlis, R. H., Doshi-Velez, F., & Gajos, K. Z. (2021). How machine-learning recommendations influence clinician treatment selections: The example of the antidepressant selection. *Translational Psychiatry*, *11*(1), 108. https://doi.org/10.1038/s41398-021-01224-x.

Kelly, T. (2010). Peer disagreement and higher-order evidence. *Disagreement*. Oxford Scholarship Online. https://doi.org/10.1093/acprof:oso/9780199226078.001.0001.

London, A. J. (2019). Artificial Intelligence and Black-Box Medical decisions: Accuracy versus Explainability. *Hastings Center Report*, *49*(1), 15–21. https://doi.org/10.1002/hast.973.

Mesko, B. (2017). Artificial Intelligence is the stethoscope of the 21st Century. *The Medical Futurist*. https://medicalfuturist.com/ibm-watson-is-the-stethoscope-of-the-21st-century/ [Accessed 1st March 2023].

Nartker, M., Zhou, Z., & Firestone, M. (2023). When will AI misclassify? Intuiting failures on natural images. *Journal of Vision*, *23*(4), 1–15. https://doi.org/10.1167/jov.23.4.4.

Nickel, F., Studier-Fischer, A., Özdemir, B., Odenthal, J., Müller, L. R., Knoedler, S., Kowalewski, K. F., Camplisson, I., Allers, M. M., Dietrich, M., Schmidt, K., Salg, G. A., Kenngott, H. G., Billeter, A. T., Gockel, I., Sagiv, C., Hadar, O. E., Gildenblat, J., Ayala, L., Seidlitz, S., Maier-Hein, L., & Müller-Stich, B. P. Optimization of anastomotic technique and gastric conduit perfusion with hyperspectral imaging and machine learning in an experimental model for minimally invasive esophagectomy. *European Journal of Surgical Oncology 2023; Apr* 18: S0748-7983(23)00444-4. https://doi.org/10.1016/j.ejso.2023.04.007.

Ontario (2022). Beta principles for the ethical use of AI and data enhanced technologies in Ontario. https://www.ontario.ca/page/beta-principles-ethical-use-ai-and-data-enhanced-technologies-ontario [Accessed 17th August 2023].

Rathkopf, C., & Heinrichs, B. (2023). Learning to live with strange error: Beyond trustworthiness in Artificial Intelligence Ethics. *Cambridge Quarterly of Healthcare Ethics*, 1–13. https://doi.org/10.1017/S0963180122000688.

Sharan, L., Romano, G., Brand, J., Kelm, H., Karck, M., De Simone, R., & Engelhardt, S. (2021). Point detection through multi-instance deep heatmap regression for sutures in endoscopy. *International Journal of Computer Assisted Radiology and Surgery*, *16*(12), 2107–2117. https://doi.org/10.1007/s11548-021-02523-w.

Strawson, G. (1994). The impossibility of Moral responsibility. *Philosophical Studies*, *75*(1–2), 5–24. https://doi.org/10.1007/BF00989879.

Subodh, S. Artificial Intelligence-The stethoscope of the 21st Century. *swatisubodh.medium.com*https://swatisubodh.medium.com/artificial-intelligence-the-stethoscope-of-the-21st-century-afdf9318c5b [Accessed 30th Mai 2023].

UNESCO Ethics of Artificial Intelligence. *www.unesco.org*https://www.unesco.org/en/artificial-intelligence/recommendation-ethics [Accessed 28th February 2024].

Ursin, F., Lindner, F., Ropinski, T., Salloch, S., & Timmermann, C. (2023). Levels of explicability for medical artificial intelligence: What do we normatively need and what can we technically reach? *Ethik in Der Medizin*, *35*, 173–199. https://doi.org/10.1007/s00481-023-00761-x.

Verma, S., Boonsanong, V., Hoang, M., Hines, K. E., Dickerson, J. P., & Shah, C. (2020). Counterfactual explanations and algorithmic recourses for machine learning: A review. *arXiv Preprint*. https://doi.org/10.48550/arXiv.2010.10596.

Wendehorst, C. (2022). Liability for Artificial Intelligence: The need to address both Safety risks and Fundamental rights Risks. In S. Voeneky, P. Kellmeyer, O. Mueller, & W. Burgard (Eds.), *The Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary perspectives* (pp. 187–209). Cambridge University Press. https://doi.org/10.1017/9781009207898.016.

Winkler, J. K., Fink, C., Toberer, F., Enk, A., Deinlein, T., Hofmann-Wellenhof, R., Thomas, L., Lallas, A., Blum, A., Stolz, W., & Haenssle, H. A. (2019). Association between Surgical skin markings in dermoscopic images and diagnostic performance of a deep learning Convolutional Neural Network for Melanoma Recognition. *JAMA Dermatology*, *155*(10), 1135–1141. https://doi.org/10.1001/jamadermatol.2019.1735.

Wirkert, S. J., Kenngott, H., Mayer, B., Mietkowski, P., Wagner, M., Sauer, P., Clancy, N. T., Elson, D. S., & Maier-Hein, L. (2016). Robust near real-time estimation of physiological parameters from megapixel multispectral images with inverse Monte Carlo and random forest regression. *International Journal of Computer Assisted Radiology and Surgery*, *11*(6), 909–917. https://doi.org/10.1007/s11548-016-1376-5.

Zednik, C. (2021). Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology*, *34*(2), 265–288. https://doi.org/10.1007/s13347-019-00382-7.

Zednik, C., & Boelsen, H. (2022). Scientific Exploration and Explainable Artificial Intelligence. *Minds & Machines*, *32*, 219–239. https://doi.org/10.1007/s11023-021-09583-6.

ZEKO (Zentrale Ethikkommission der Bundesärztekammer). (2021). Entscheidungsunterstützung ärztlicher Tätigkeit Durch Künstliche Intelligenz. *Deutsches Ärzteblatt*, *118*, 33–34. https://doi.org/10.3238/arztebl.zeko_sn_cdss_2021. A1-13.

Zhou, Z., & Firestone, C. (2019). Humans can decipher adversarial images. *Nature Communications*, *10*(1334). https://doi.org/10.1038/s41467-019-08931-6.