AI and Ethics (2023) 3:567–584
https://doi.org/10.1007/s43681-022-00189-x

ORIGINAL RESEARCH

Are we justified attributing a mistake in diagnosis to an AI diagnostic system?

Dina Babushkina[1]

Received: 16 March 2022 / Accepted: 2 June 2022 / Published online: 9 August 2022

Abstract

Responsible professional use of AI implies the readiness to respond to and address—in ethically appropriate manner—harm that may be associated with such use. This presupposes the ownership of mistakes. In this paper, I ask if a mistake in AI-enhanced decision making—such as AI-aided medical diagnosis—can be attributed to the AI system itself, and answer this question negatively. I will explore two options. If AI systems are merely tools, then we are never justified to attribute mistakes to them, because their failing does not meet rational constraints on being mistaken. If, for the sake of the argument, we assume that AI systems are not (mere) tools, then we are faced with certain challenges. The first is the burden to explain what this more-than-a-tool role of an AI system is, and to establish justificatory reasons for the AI system to be considered as such. The second is to prove that medical diagnosis can be reduced to the calculations by AI system without any significant loss to the purpose and quality of the diagnosis as a procedure. I will conclude that the problem of the ownership of mistakes in hybrid decision making necessitates new forms of epistemic responsibilities.

Keywords Automated diagnosis · Medical AI · Ethics and epistemology of AI · Hybrid decision making · Responsibility for diagnosis · Ownership of mistakes

1 Starting point

This paper is a response to the strengthening voice in favor of diagnosis automated through AI technology. This tendency can be seen both in the research literature and in a growing number of companies offering AI diagnostic software. Such diagnosis is envisioned to be carried out either solely by specially designed AI programs or by a sort of hybrid agent (AI-system + human), in which a significant part of cognitive workload is performed by the software. What I would like to focus on specifically in this paper is the tendency to limit or even push human expertise out of the diagnostic process or otherwise de-professionalize certain aspects of medical diagnostic procedures.

One of the core ethical concerns about automated diagnosis is the responsible[1] use of AI diagnostic systems in

✉ Dina Babushkina
d.babushkina@utwente.nl

[1] Faculty of Behavioral, Management and Social Sciences, Section of Philosophy, University of Twente, Enschede, The Netherlands

[1] To clarify the scope of the paper: it will only discuss the ethically relevant dimension of responsibility, omitting the analysis of its legal aspects which is a research area in its own right. As a result, the reader should not expect a discussion of the legal consequences of medical mistakes, or an analysis of responsibility attribution to AI as a legal entity. For the legal implications of the use of AI in healthcare, the reader may consult, e.g., Gerke et al. [27], Schneeberger (2020), Schönberger [48].

health care (e.g., [29, 30, 36, 41, 46, 49, 54, 56].[2] But what constitutes a (morally) responsible[3] professional use of such systems in a high-risk context of human health and well-being? And more specifically: what constitutes a responsible incorporation of AI systems into the diagnostic decision making process of a medical professional? These are not easy questions to answer, especially since answering them requires understanding which demands are rationally justified, and therefore must be met, and which demands may be dismissed in the view of the natural process of evaluation of the diagnostic process.

Philosophers usually approach the problem of responsibility in connection to artificial agents from a general metaphysical standpoint. This presupposes investigating the concept of agency, by either trying to estimate whether artificial agents (AA) such as AI systems can fulfill the necessary and sufficient conditions for being moral agents, or to re-configure the concept of the moral agency in such a way that it would allow one to attribute responsibility to AAs at least *in some sense*. This debate is extensive and well summarized in the majority of articles on the topic of responsible artificial agents. To avoid unnecessary repetition, I refer the reader to Gogoshin [28], Hakli and Mäkelä [31] or Behdadi and Munthe [6] for comprehensive overviews of the debate on the question of responsibility by artificial agents.[4] In this paper, I will employ a somewhat different strategy. I will approach the problem of responsible professional use of AI-based diagnostic tools via a less discussed problem of *the ownership of diagnostic mistakes*. My main concern will be: Are we justified in attributing a mistake in diagnosis to an AI-based diagnostic system?

2 Responsible professional use of AI-based diagnostic tools: problem formulated broadly

For the purposes of this paper, we can understand responsibility as a way of dealing with harm resulting from one's actions. In application to AI tools, this translates into a certain manner of addressing potential or actual harm as a result of decisions (and actions) based on the AI outcome. What manners of dealing with such harm amounts to a responsible stance? It is reasonable to assume—from the ethical point of view—that this includes at least the following commitments on the part of the agent: the prevention of unnecessary harm,[5] the mitigation of unavoidable harm, appropriately addressing the situations when harm has occurred (i.e. to prevent it from re-occurring, repairing the damage), properly acknowledging the avoidable harm in decision making (e.g., by exploring alternative options), acknowledging the wrong. Such commitments characterize the responsible stance of the agent. On the other side—especially in dependable relationships such as the relationship between the patient and a medical professional—there are certain expectations that the one (potentially) affected by the agent's actions is rationally justified in having. These expectations typically concern the actions of a medical professional, but they also extend to knowledge, affective attitudes, and professional commitments. In application to patient–doctor relationships this implies that there are certain things patients are rationally justified in expecting a competent professional to know and do.

It is rational, for example, to expect that a responsible medical practice is guided by the imperative of harm avoidance and preventability: avoiding unnecessary harm and preventing the possibility of a patent coming to harm as a result of medical interventions. In other words, if it is possible for a medical professional to acquire knowledge about preventing an injury by a tool or procedure used on a patient, this knowledge should be applied to prevent such an injury. One example is targeting an X-ray beam only on a certain part of the body, while the rest of the body is shielded from the radiation.[6] For the same reason, medications are accompanied by instructions of use, for example, concerning the maximum dose, which, if exceeded, would be harmful. In both cases, creating ways to prevent possible harm (safety guidelines) required understanding how the tool/medication works and research on the limitations of their applicability. Being knowledgeable of such limitation is a part of competence. AI diagnostic tools are not exempt from his principle. Furthermore, from the view of a patient's autonomy (as well as her ability—or that of a third party—to evaluate her health situation), it is equally reasonable to expect that the

[2] For an overview of ethical questions about AI in healthcare see, e.g., Trocin et al. [56], Burr et al. [11], Morley et al. [40], Stahl and Coeckelbergh [50].

[3] On the difference between moral and non-moral responsibility see, e.g., Tigard [54] and [4].

[4] On responsibility gap with respect to AI see, de Sio and Mecacci [19] for typologies, and Tigard [55] for an argument against.

[5] Sand et al. [46] propose an account of forward-looking responsibility for AI in health care, by which they understand moral requirements to prevent harm from happening.

[6] Please note that I do not suggest an analogy between the effects of the radiation or pharmaceutical drugs on the body of the patient and the work of AI on images; these are offered merely as examples of possible harm prevention (regardless to the type of harm) that are practiced in medical domain.

medical professional would be committed to transparency[7] with respect to ways and methods she reaches the conclusion concerning the patient. And again, AI tools are not an exception. One reason why transparency is important is its connection to the ability to give an account[8] (cf. [10] of one's decisions and actions (inc. the explanation of relevant reasons and factors), especially—but not only—in the situations when the patient has been harmed.

The readiness to respond to and address—in an ethically appropriate manner—the wrong that may be a result of one's own actions (and, in the case of medical diagnosis, with the use of certain tools/procedures) is an essential element of a responsible stance toward one's professional action. This means that neither the harm that has already happened nor the possibility of things going wrong may be ignored, be made to appear less important (or as non-preventable), be discarded as accidents (provided, of course, that they are not genuine accidents). In other words, responsibility presupposes the ownership of mistakes,[9] by which I understand.

> an appropriate attribution of the harmful event to the action/omission or misapplication of knowledge by an agent, on the basis of it being her obligation due to professional competence.

This applies to mistakes in diagnosis. Inability to avoid preventable mistakes maybe an indicator of professional incompetence (when external factors are excluded). By preventable mistakes[10] in the context of diagnosis I mean such situations when it would be reasonable to expect that any properly trained medical professional could have made the correct diagnosis, but the misdiagnosis has occurred, nonetheless. Under this clause, the failure to diagnose correctly entails a mistake in the process of diagnosing and this mistake has to be appropriately attributed and addressed.

3 Attribution of a mistake in misdiagnosis: narrowing the problem down

The problem with a diagnosis that is carried out by a hybrid agent (or by the AI system alone) is that it is unclear who/what is to be attributed the ownership of the mistake in the case of a wrong diagnosis. There are different ways this can be dealt with. From the legal point of view, there might be an inclination to limit the question to that of the retribution and punishment, and start exploring for possibilities of attributing the mistake to AI systems alone (e.g., by making them legal persons), or to medical administration, responsible for the decision to incorporate the AI system in the medical center, or, in a sense to no one, through a certain type of insurance system, or perhaps, a combination of the above or above and the medical stuff. From the perspective of Dignum's [20] framework for Responsible Research and Innovation, one would be inclined to look at various societal actors (such as: researchers, developers, businesses) involved in the development, marketing, and implementation of AI-diagnostic tools in society, to explore their potential role in the mistake that has occurred as well as their contribution to the explainability—or the difficulty thereof—of these tools that, in turn, contribute to the transparency of the decision that AI augments or supplements. Each of the approaches has their internal reasoning and would require an independent investigation. In this paper, however, the concern is not so much with the retribution or punishment, but with much broader spectrum of actions and attitudes toward harm. One of the main concerns here is *offloading* the responsibility for misdiagnosis to the AI system itself which, if it takes place, undermines—if not eliminates—this broad spectrum of actions and attitudes.[11]

Offloading responsibility is ethically problematic because it undermines the possibility of blame where it is due and allows for blame where it is not justified. In other words, it removes from the scene the agent, to whom it is appropriate to assign the responsibility for harm and introduces *a moral substitute*, i.e. an agent that acts as an object of blame (and retribution) without being the subject of wrongdoing. There is a high risk that offloading responsibility for misdiagnosis to AI-tools may happen when human agents over-rely (cf. [39] on the algorithms without properly assessing the limitations of their applicability or do not have

[7] In a boarder sense, such commitments could be seen as falling under the category of professional duty to care.

[8] Accountability, when understood as an obligation to explain one's actions and decisions [17] is an important element of responsibility. However, the terms responsibility, accountability, and liability are often used synonymously, and their meaning may differ depending on the context. On accountability with respect to the use of AI in healthcare see, e.g., Habli et al. [30].

[9] The concept of ownership of mistakes is introduced here from the ethical rather than legal point of view. The intention behind this concept is to, on the one hand, tie professional decisions and actions of a person to her moral agency, and, on the other hand, to show what follows from this connection. As a result, the discussion of legal agency as well as the question of legal responsibility and liability are outside of the scope of this paper. As suggested by a reviewer, one interesting question is the effect that a specialist's degree of competence (due to his/her career stage or specialization) may have on legal grounds for mistake attribution.

[10] Since the ultimate concern here is moral responsibility, I am talking about non-trivial mistakes, i.e. those that amount to failure to diagnose correctly (not all mistakes in the diagnostic process necessarily lead to the failure to diagnose).

[11] In this respect, Tigard [53] is interesting as he argues for the preservation of blame taking in medical practice.

a proper understanding of the meaning of algorithmic output. The strong push toward automated diagnosis when AI system alone is tasked with a judgment about the patient's condition, excludes (or otherwise significantly limits) the expert participation and judgment in the diagnostic process, understandably alienating the human agents from the decision making and, as a result, from the *sense* of responsibility. This creates favorable conditions for transferring responsibility for possible mistakes to the AI-systems since they are, at least *de dicto,*[12] presented as the agents carrying the diagnosis. To some extent, we could try to mitigate this danger through the category distinction between what can be called an assistive AI and replacive AI. This would require a systematic philosophical analysis of each of these categories, as well as the understanding to what extent—if at all—and in which contexts, a replacive AI is permissible. However, there is still a risk that even when this theoretical distinction is clear, on the practical level we will face moral sloppiness. A decision-maker, even when informed about the assistive role of the AI system, may still be tempted to remove herself from the responsibility for the decision, for example, because of a belief in epistemic authority of the algorithm.[13]

In what follows I will focus on the question: *Are we justified attributing mistakes in diagnosis to the AI system that has been employed in the diagnostic process*?[14] If not, offloading the responsibility will lead to blame vacuum, i.e. the situation where the AI system is playing the role of the offender but cannot be an appropriate object of blame [4]. In what follows, I will argue that we are not justified in making such an attribution. I will first investigate if we are justified in ascribing an act of mistake to a piece of software and whether an AI system capable of mistakes at all. In approaching these questions, I will take an ontological route. I will argue that we have two options: we either (1) see AI systems as mere tools, or (2) we do not. Turning to the first option (1) first, I will show that if AI systems are indeed tools, then in the context of the diagnostic process, the use of AI systems amounts to no more than a use of a sophisticated tool by a medical professional (akin to laboratory tests). In this case, we are not justified in attributing diagnostic mistakes to a tool because a tool's failing to produce the correct result does not satisfy the normative constraints on mistake attribution. Instead, we are only justified to attribute errors to them. Next, I will consider the second option (2). I will show that if one, for the sake of the argument, assumes that AI systems are, in some sense, capable of mistakes, one faces certain challenges. The first is the burden to explain what this more-than-a-tool role of an AI system is, and to establish justificatory reasons for the AI system to be considered as such. The second challenge is to prove that medical diagnosis can be reduced to the calculations by AI system without any significant loss to the purpose of the diagnosis as a procedure. My goal will be to show that there is a high risk of such losses, especially from the perspective of the patient. I will invite readers to explore and compare two scenarios. One will describe (a) what the diagnostic procedure would look like if it were reduced to the epistemic processes of an AI system based on deep neural networks, and (b) what this would entail for the patient and the reality of her experiences of the sickness. This scenario will be contrasted with another, in which the diagnostic process—together with the criteria of success/failure—follow "the ideal path" which consist of the best diagnostic practices as they are described in the relevant literature.

4 If AI-systems are a tool

If AI systems are merely instrumental to the diagnostic process, the question is whether it is sensible to attribute misdiagnosis to a tool. To understand when we are justified to attribute a mistake to an entity, we need to know the revenant constraints on a mistake attribution. First thing to note is that mistake is different from error, and register an error is not the same thing as to attribute a mistake. Error is a concept that grasps the *discrepancy between the actual result and the expected result/truth*. It is a form of deviation, which that must be corrected to "hit the mark". Mistake, on the other hand, is a concept that grasps *the connection between the failure to X and the agent's cognitive states*. This connection is subject to, at least, one epistemic and two rational constraints. *The epistemic constrain* can be spelled out the following way:

> (1) An agent can be said to be mistaken about X iff the agent is aiming at the correct judgment (or belief) about X and fails.

[12] Here *de dicto* means "by expression only", "on words", "as portrayed by various narratives, incl. marketing". *De dicto* is opposed to *de re* which means "according to facts", "as things are".

[13] This is a reason why, in this paper, the question of attributing responsibility to an AI system is considered in a manner that would accommodate both cases.

[14] I will not attempt here to answer who is to own a mistake in the case of a wrong diagnosis when it has been carried out by a collective agent. This is a complex issue, requiring an analysis of different types of situations and evaluation of the degree of involvement of various parties, such as medical personnel, medical administration of the center where the diagnosis has been carried out, as well as the system's developers. My concern is narrower and more specific—it is a worry that the ease with which the claims that AI technologies pave way to doctor-less diagnosis makes is equally easy to offload the responsibility for the misdiagnosis to an AI system itself.

It is easy to explain this conditional on an example when an agent holds a belief contradicting facts. Let's say A believes that M stole his wallet. However, the truth is that A forgot the wallet at home. For us to consider A's claim that M stole his wallet a mistake we must assume that A aimed at being correct even though he failed (i.e. his claim was a consequence of a sincere belief that M stole the wallet). If, A knew he was wrong, then he would have been lying. If, on the other hand, he had no interest in knowing the truth, A would be bullshitting [26] or trying to manipulate the situation in some way. Were B to know the facts and point to M's mistake, B would be able to reach to the subject whose cognitive states were the reason for the failure to appropriately judge the situation and, as a consequence M would be able to correct his judgment in the face of the fact as well as to clear M from the accusations.

Whether we are talking about beliefs alone or also about actions, being mistaken is failing to be correct/right due an inadequacy of one's knowledge or its misuse. For example, a mistake may occur due an incomplete knowledge: one may not take into consideration all relevant circumstances and made a wrong conclusion on a partial basis. Ignoring or not being aware of symptoms relevant to the identification of a condition is one example. Alternatively, the problem might be in the misapplication of a piece of knowledge to the situation at hand, such as explaining the phenomenon in terms of irrelevant regularities. Yet another example of a mistake is *fallacy* i.e. an error in reasoning due to invalid logical forms of inference. An example would be a concluding that something is the case because it is possible (or might happen), or deriving a conclusion about the condition/behavior of an individual X based on the premise about what it typical to some X.

Now, if we turn to the demands of rationality on mistake attribution, *the first rational constraint* can be formulated this way:

> (2.a) An agent is only able to make a mistake if it is *possible* for the agent to deviate from the resolution path which guarantees the correct result.

This is a matter of logical necessity. It is a condition that makes a mistake possible. Let's say, A has a task to solve the problem X. There is the right procedure that can be followed to guarantee the correct result for X. If the only thing that A can do is to follow this procedure, then A will always be right. It is much like being a hamster in a wheel: as long as the hamster can only run inside the wheel, it cannot deviate left or right. Being mistaken is a lot like sidetracking: an option of not doing it right must be available to you. When the condition 2.a. is not met, but the result happens to be wrong, the *agent is not at fault*. The cause of error should be looked for elsewhere: perhaps, a third party intervened and altered the conditions, rendering the fixed resolution path inapplicable.

I will formulate *the second rational condition* this way:

> (2.b) One may attribute a failure to a mistake on the agent's part only if the agent was *at liberty* to deviate from the predetermined resolution path.[15]

This is a condition that a failure has to meet to qualify as a mistake on the agent's part. This is also a matter of rationality of mistake attribution: it only then makes sense to attribute an active role in a failure to an agent (such as a failure to produce an expected result or the correct solution to a problem) when this failure has resulted from actions, beliefs, and decisions that were under the agent's control. This is significant because attributing mistake to someone is an act that sanctions certain practices, such as blaming, correcting the erroneous belief (e.g., to prevent the mistake happening in future), or (possible) punishment. As a result of 2.b., if the failure has originated from the agent's cognitive states but the agent was not at liberty to deviate from

[15] A reviewer pointed to an interesting area of discussion in connection to this condition: the extent to which the tendency of automation bias may affect a human decision-maker's (such as a welfare worker, as discussed in [44] liberty to deviate from the predetermined resolution path (coded in an algorithm that she is provided to carry out her work). Even though I won't be able to discuss this multi-dimensional issue in length in this paper, I would like to offer a brief comment. One way we could approach this issue is from the perspective of the ethics of positional duties (i.e. duties that one has in virtue of occupying/having a certain social position/role). In this case, one could say: because the decision-maker carries out her duties in a pre-determined matter (i.e. in a matter prescribed to her by a superior and as a work requirement), we cannot attributed her a mistake that has resulted from her carrying out the prescribed procedure. However, this does not stand, because, even as a person who is subject to a positional duty (in this case, the use of an algorithm), she is not free from the moral duty (i.e. the duty that she has as a moral being) to do what she can to prevent (or minimize) harm that may result from her decisions. From the moral point of view, to be bound by a social requirement is not the same as to have no liberty to deviate. The liberty to deviate is about conditions to exercise one's free will. So, if the use of the algorithm by a human results in a wrong decision, then she has her share in the mistake, unless the conditions under which she could exercise her free will were suppressed. The extent of her share, i.e. her exact contribution, so to say, to the mistake – is subject to further analysis. But in any case, as a moral agent, the decision-maker is capable of acting/not acting upon reasons, capable of knowing the moral properties of her actions and evaluating means/tools by which she reaches her decisions. A few things follow from this. The latter condition is a foundation for the epistemic responsibility of the decision-maker: i.e. the moral responsibility to understand how the result of the algorithm you use relates to knowledge (e.g. it limitations, biases, etc.). So, if the decision-maker finds herself being required to use a certain system, then she has a moral obligation to know, among other things, how it produces its output and what its limitations are. At the same time, she has a moral right to refuse relying on the systems if she has a reason to believe that it is biased and may result in unfair treatment of others.

them, *the agent can be excused*.[16] This would be the case if A's belief that M stole his wallet was a result of a hypnosis, and A was not aware of this idea being implanted in his head and could do nothing to prevent it.

In the case of tools, both epistemic and rational constraints fail to apply. On the one hand, tools are not at liberty to deviate from their function or the solution path. In fact, the predictability is one of the main criteria of a good tool. And then tools (narrow sense of a tool, excluding the possibility of considering a human person as a tool) are incapable of epistemic attitudes and therefore do not aim at truth/being correct. Hence, tools do not make mistakes. This does not mean that tools cannot produce a wrong result. Tools err and lead to mistakes by humans who rely on them. And the possibility of tools producing a wrong result with a subsequent influence on the human decision-makers must not be taken lightly, especially in the case of AI. When it comes to the severity and scale of consequences of mistakes due to the influence of AI in decision making, AI stands apart from other tools for at least the following reasons. One is the notorious difficulty of explaining how the system reaches its outcome—what is often referred to as the "black box" problem. This only becomes more serious if we take into account, one the hand, the capacity of DNN to process enormous amount of data, and, on the other, the fact that models come tuned to produce highly reliable, but still contextual results where understanding and defining the parameters of this contextuality present yet another epistemic problem. The second reason is the fact that the use of AI leads to automation of decision making. Due to automation, the decision making becomes insensitive to specific details of various cases/situation the models are applied to. This often means that if a case does not fit the model, so much worse for the case. But also, again, given the ability of DNN to process enormous data in a short amount of time, if an AI systems produces a wrong or unfair outcome, it can do so on an enormous scale in a short time too. All this significantly raises the moral cost of mistakes.

Since it does not make sense to ascribe a mistake rather than an error to a tool, the tool's failing should be explained differently than making mistakes. There are a number of such explanations possible. For example, the tool might produce a wrong result because it has been misapplied or calibrated imprecisely by the user. In this case, we are dealing with a mistake by the user of the tool while there is nothing wrong with the tool itself. A bug in a program due to the programmer's mistake in syntax, data type or function request, in appropriate model architecture or activation function, a trained AI model applied on wrong type of data—all these cases are misapplication on the side of the user. In the case of a DNN model, it might be applied to cases which are not representative of the dataset it has been "trained" upon. Another explanatory category can be reserved to the cases when one is dealing with a bad/faulty tool which is inaccurate or broken. In this case, it is more appropriate to describe the tool's failing as an error. Here are some examples: the error in measurement in a digital blood pressure monitor due to low battery, or erroneous results due to a malfunction in the computer system. In the case of AI, we made describe this way a model, in the creation of which the programmer did not follow proper validation techniques.

5 What if AI systems are "something more" than tools?

Given the ontological status of AI diagnostic systems—i.e. these systems being a type of technology among other types of technologies[17] and technologies do not *by default* enjoy any privileged status as, for example, persons do—it is not self-evident that such systems are to be treated as "more than mere tools" when it comes to their capacity of owning mistakes. Thus, those who claim that we should do so—and if they wish this to go beyond a mere claim—bear the burden of explaining what this role can be and justifying it taking into considerations, on the one hand, the purpose and inner logic of diagnosis (e.g., we must make sure that treating AI as more than a tool would not undermine diagnosis and its goals), and, on the other hand, the reality of what we are warranted to expect from AI, given its epistemological and scientific limitations.

Let us, however, for the sake of the argument assume that AI systems can be seen as entities that are, *in some way*, capable of mistakes. Those who would want to defend this, would have to first explain what these entities could be. But even if these entities were persons much like human persons, we would still not be justified to attribute a mistake in diagnosis to them. Were we to do so, we would reduce the success/failure in diagnosis to the success/failure of an AI-system to fulfill its task. It is, however, questionable that such reduction is justified. To show why it is so, in what follows I will invite readers to imagine a scenario where such reduction is made a norm for any diagnostician and to explore what this entails for the process of diagnosis and its goals. The idea is, on one hand, to unwrap a possible future of diagnosing, the reality that it would create and the way this would affect the patient herself. On the other hand, the idea is to engage in the dialog about the desirability of

[16] For more on the philosophy of excuse, I refer the reader to [2] and [59].

[17] This is only to say that AI systems are tools from the ontological point of view, but this does not mean that AI as tools are not distinct from other tools.

such scenario, given the patient's perspective. For this, I will invite the reader to contrast this scenario with another one, where no such reduction is made. This second scenario will be based on the best diagnostic practices as they are described in specialized literature. The ways one can succeed or fail in reaching the diagnostic goals in this scenario are radically different from the first one. Contrasting these two scenarios, I intend to show that that such a reduction is open to a serious objection as it entails an unjustified loss of important elements of diagnostic practice. Some of these elements aim to minimize the possibility of mistakes associated with insufficient contextualization of medical knowledge in the physical reality of the patient's and her experience of the symptoms.

6 Agent Π and agent Ω

Imagine two scenarios. One has a potential to create a world[18] which is very much like our own. Here the diagnostic process is carried out by an agent Ω, very much like diagnosticians in own world, but in this possible world the diagnosis is always carried out in *the ideal way,* i.e. according to best practices as they are described in specialized literature. The second scenario will lead to a very different world. Here the task of diagnosing patients is fully transferred to diagnosticians such as agent Π, who performs the same cognitive role as an AI system based on deep neural networks (DNN). The agent Π is identical to the AI system when it comes to functionality, abilities, tasks, and the method accomplishing the tasks, but—she is, in fact, a human.[19] We could assume that in this possible world, people can modify their brain in a way that allow them to perform the same type of computations as algorithms, in exactly the same manner. However, transforming the human cognition into machine processes, comes with the limitations of machine epistemology: like an AI-system Π-diagnosticians cannot deviate from the predefined algorithmic decision process.

Let's call the methods that agents Ω and Π use to fulfill their diagnostic role ω-ing and π-ing, respectively. ω-ing would then be based on the internal logic of the diagnostic process considered standard in pre-AI era. By contrast, π-ing would be full equivalent of an AI model based on deep neural networks. And finally let's say there is T, a multiverse traveler who is able to move between the alternative scenarios. And so, T gets sick, experiencing fatigue, chest pain and difficulty of breathing, and is trying to find out what is wrong with him by first getting examined in the Π world and then in the Ω world. Π, based on the method π available to her, makes a judgment about T's condition, but it turns out to be incorrect. The question now is: *Can the mistake in diagnosis be reduced to the mistake in Π's judgement?* Or rather: Are we justified in reducing the failure to diagnose (correctly) to the failure in pattern recognition, given its limitations? Or would such a reduction entail a loss of something significant for the diagnostic process?

To make the explanations of notoriously complex DNN easier, as well as to avoid getting lost in the details of possible discrepancies between existing diagnostic models, I will use imaginary simplified example which does not necessarily translate into any specific case described in literature. However, if the reader is curious what such models might look like, I refer him/her to Irvin et al. [33] and Rajpurkar et al. [43]. Please note that these are mentioned only as an illustration, and I do not investigate any specific claims that either group of researchers are making. Instead, I am looking into a hypothetical situation when one might be tempted to treat an output of AI system as equivalent to the judgment about a diagnosis. For the purposes of my argument, it does not matter whether this temptation presents itself merely as discourse tendencies fueled by hype narratives around medical AI, or whether it reflects actual practices where an agent finds herself inclined to make a cognitive jump from a thought: "AI system indicates X" to "the AI's output means mostly likely X" and to "therefore T has X" (or, "therefore T has disease D which is characteristic for X).

7 The world of Π

The agent Π works with T's case in the following set-up. Π understands her task as comparing the image she is given (which *we* know to be an X-ray of T's lungs) with a reference set of patterns, represented by labels. Each label carries the name of a lung condition (let's refer to them as "condition x", "condition y", "condition w"). The agent Π has no knowledge of how the set of patterns is compiled. She has no way of checking whether the labels correctly identify conditions they refer to (labels are created by the agent Λ who does not share her labeling criteria with Π). Furthermore, if we wish to follow the analogy as close as possible, Π would not even know that the image in front of her is an X-ray which depicts lungs. For her, this image is just a carrier of information about the intensity of colors at each point of the image. Π will be determining T's condition without ever

[18] It is worth noting that this is not meant to be to a "possible world" argument as it is typically used in analytic philosophy. In this paper, I refer to two worlds as possible scenarios to illustrate a problem with AI in automated diagnosis, explore the consequences of certain assumptions, and to make the complex discussion of the epistemology of DNN more relevant and easier to grasp.

[19] The idea is to show that what is at stake is not so much whether the diagnostician is a human or machine, but the sort of cognitive operations that are involved in the process of diagnosis.

seeing the patient. Nor will she ever know anything about T's symptoms, the way they developed, or T's underlying conditions. Agent Π works alone, does not discuss his conclusions with other diagnosticians.[20] Neither is she at liberty to use any other method that the one she is prescribed, π.

8 What would it mean to succeed in π-ing?

π-ing, as the only method available to Π, consists in pattern recognition. This may sound straightforward, but it is far from that. Pattern recognition, even though currently mostly associated with computer science, is not unique to it. It is an epistemological category which refers to certain ways in which mind generates knowledge. What often causes confusion is the fact that pattern recognition is an umbrella term for different types of epistemic tasks aiming at establishing similarities and regularities. Depending on what sort of similarity/regularity the agent aims to establish between things she observes, she might find herself performing different epistemic procedures, each qualifying to be called patter recognition. In other words: it matters what you do while looking for a pattern and why you do it. Confusing these different procedures will result in misplaced expectations from the algorithm and is bound to lead to misinterpretation of the results.

Now, what would count as Π's failure to π, depends on what would count as success. Different epistemic tasks underlying this or that case of pattern recognition have their own criteria of success. More specifically, π as a pattern recognition process *could* be understood as consisting in:

(a) Establishing *a* similarity, that is to say, drawing some sort of connection between observations (such as T's X-ray image and X-ray images from other patients), finding *any* type of similarity. There is no specific requirement about what this connection should be; it may turn out to be relevant or irrelevant to the question one is trying to answer.

(b) Selecting a combination of elements in an observation (such as T's X-ray image) that would connect it to a pre-established pattern(s), i.e. those patterns in virtue of which different images of a dataset are considered similar under such labels as "condition x", "condition y", "condition w". And again, this connection does not have to be of any specific kind; that is to say, it may connect the observations trivially (not discriminating between any bits of data and using such epistemological noise, as background features, empty spaces, imperfections in the image, for drawing similarities) or it may capture the essential link between the observations [5]. There is just no way of knowing.[21]

(c) Indirectly, in a somewhat trivial sense, pattern recognition may be used to refer to a combination of different cognitive tasks (or varied levels of complexity) that altogether may be better called solving a problem,[22] as in answering the patient's question of the sort "What is wrong with me?" This amounts to a rich sense of "identifying the disease", which includes, for example, such cognitive operations as finding out whether the condition that is observed meets necessary and sufficient conditions of a disease X.

Now, this being so, if one wished to argue that agent Π could be considered as owning the diagnostic mistake—when the diagnosis is understood in this rich sense of answering the patient's question: "What is wrong with me?"[23]—one would have to assume that her task is something like (c). This assumption, however, is very hard to substantiate given the proper understanding of what that it

[20] A reviewer offered a possible objection to this part of my analogy. The objection goes something like this: despite the evident goal to highlight the lack of collegial shared decision making, the analogy does not seem to be precise because the AI system—as opposed to a human—is introduced to massive datasets of X-ray images. These are gathered from thousands of patients. Since each image—according to the best practices of "training" a model—comes with a label that connects to a confirmed diagnosis, the AI system can be said to build upon the collective experience of thousands of medical professionals. Arguably, this puts the AI system in a more advantageous position, because the human diagnostician would rely only on her (considerably narrower) experience with such images (X-rays). This is an interesting objection to explore in detail, especially from the epistemological point of view. However, due to limited space, I will only note that that the crucial difference here is the inaccessibility of dialog and critical revision (including the exchange of reasons) that is characteristic to the collegial shared decision making. From this point of view, it is more accurate to say that Π is exposed to a compendium of resolved cases, rather than to say that she is participating in a medical concilium.

[21] This is a part of what makes DNN a "black box". The difficulty to explain the output of DNN has been extensively discussed in scientific literature. In application to (non)-explainability of A in I healthcare context, I recommend Durán & Jongsma [22].

[22] In this respect, the way Stanley & Campos describe a difficult case of a patient with a complex clinical picture is especially relevant: "So begins the process of diagnosis based upon the signs, symptoms, laboratory testing and imaging, the family and personal history. Diagnosis involves surveying the horizon for hypotheses that identify causes for episodes of pain, nausea, and diarrhea. Such hypotheses are generated through a process of abduction. Which sign, symptom, laboratory test or image—or combination of these—will provide a fruitful entry to solving this problem?" (2013, p. 303).

[23] Cf. "What is wrong with me?" (cf., e.g., [35], 14) is a request to solve the problem as it presents itself to the patient. It is a generic formula that reflects the patient's interests within the diagnostic process. These can be different from the narrow medical interest; they render "accurate diagnosis valuable even when this diagnostic is not clinically effective" [35], 36). Answering the patient's question "What is wrong with me?" presupposes an exploratory search and, perhaps, even "thinking outside the box", especially when the case is complex.

Table 1 A schematic representation of π-style decision making about the patient's condition

	1. Detailed pictorial analysis of the image at hand (which we know to be an X ray of T's lunges)
Contextualization in the context of the set of patters calculated by K and labeled by Λ	2. Calculation of the degree of relevance of this image to a predetermined set of patterns
	3. A judgement about the patient T's condition (which will further be used for the purposes of treatment)

is that a DNN actually is doing when it is carrying out its pattern recognition task. The truth is, even though often hoped to, DNNs do not perform such high order open-ended problem solving. Assuming (c) reveals a confusion/conflict between the internal and external criterion of success/failure. The external criterion comes from the comparison of the Π's result (= the output of AI-system) with the goals of the external observer and what she expects from Π. The internal criterion is established by comparing the result of the Π's comparison (= the output of AI-system) and the goal as is exists *for* the algorithm. These two do not necessarily coincide. It is the internal criterion—which reflects the actual task that the agent is carrying out—serves as a rationality check on the justifiability of the external criteria. Let's try to reconstruct, in a schematic may, what is involved in π-ing by the analogy with image processing DNN.

9 π-ing and decision-making

Π carefully scans the X-ray she is given, millimeter by millimeter, and calculates the color values of each segment (from 0 to 1). She then averages, in a predetermined manner, these values for bigger sections of the image. As a result, she produces a map of averaged weight distribution of color-density. She then compares this map with a set of patterns she has been provided for referencing. By comparing the map to this set of patterns, the agent Π is able to make a technical statement about the degree to which her map is similar to each of the patters matching each label. Something like this: "T's image is 10% similar to the pattern attached to the label "condition x", 40% to "condition y", 50% to "condition w"" (this makes Π's work analogous to so called "prediction stage"[24] of a DNN model).

When carrying out the comparison, Π is using a special tool, *a comparison filter*. This tool is created by the technician K (analogous to so called "training stage"of the DNN model). Π does not have any influence on K's work, neither is she able to assess her results. The agent K prepares the comparison filter by analyzing color distribution among a large dataset of images, the content of which is unknown to her (but which *we* know to be X-rays of various lungs conditions). The result is something like a lens through which Π is able to review the color-density map, filter out certain types of features and determine how similar the image is to each of the provided labeled categories. There is a calibration filer for each category, but the parameters of the filter are hidden in the calibration tool itself and are not accessible for Π to assess. She receives the tool prepackaged, together with the reference pattern. Π's task is to evaluate numerically ("weigh") a degree to which each feature (related to color density) in the image in front of her (and which we know to be T's X-ray) is represented by the pattern. In other words, she essentially *projects* (or reads) into the image a combination of elements which has been pre-calculated by K as typical for each labeled group of images belonging to K's reference dataset. So, basically, Π's technical statement should be more accurately understood as: "T's image is 10% similar to the pattern calculated by K for images gathered by a labeler Λ under the label "condition x", 40% to "condition y", 50% to "condition w"".

And so, schematically, Π's decision making process about the patient's condition would look something like Table 1.

10 What would it mean to fail in π-ing?

Let's now go back to the list of possible ways one could succeed in pattern recognition and see whether any of them express warranted expectation from Π, with respect to which we can evaluate his performance (given the nature and the limitations of DNN from an epistemological point of view). It is here that we should search for the internal criterion of failure. Let's start with first two alternatives. Depending on whether which of the

[24] Babushkina [5] argues that we should move away from inaccurate terminology describing various stages/elements of DNN models. According to the researchers, it is more appropriate to refer to this stage as "pattern project stage", while what is commonly called "training stage" is better described as "pattern generation stage".

two meanings of pattern recognition we refer to, Π can be legitimately accused in failing π-ing in either of two ways:

(a) Failing to find/establish a(ny) pattern at all. In this case, the agent Π did not succeed in finding any similarities between the T's X-ray and the images included in the reference-database.

(b) Failing to select a combination of elements in an observation (such as T's X-ray image) that would connect it to a pre-established pattern(s) (e.g., those patterns in virtue of which different images of a dataset are considered similar under such labels as "condition x", "condition y", "condition w".

Strictly speaking, Π has no access to the reference database. For this reason alone, it does not make sense to ascribe the first type of failing, (a), to her as well. But I would argue that even if we break the analogy with how the DNN algorithms typically work and—for the sake of the argument—stipulate that Π can be trained in K's work, it would still be impossible for her to fail in the sense described under (a). So, let's assume that Π calculates the comparison filter herself. This would mean that Π has direct access to the reference database of images of various lung conditions and as well as to the labels (keep in mind that even in this case, Π would not have access to the labeling criteria). This would invalidate the argument in the beginning of this paragraph. But the truth is that Π will always find *some* similarity; however trivial it might be. This is because π-ing does not discriminate between essential and non-essential features (such as the background upon which the lungs are displayed). Depending on the level of abstraction, everything is similar to everything. This feature of DNN has been widely explored in AI art,[25] where the output of a model is a combinatory image, reading feature of one image into another (say, a Van Gogh's *The Starry Night* into a cityscape, or that of an image of a fish into that of a dog). So, we must dismiss (a).

(b) is a more likely way to fail which is open to Π. Let's remember that the diagnostician Π has a very narrow function, the only thing she can do is to compare T's X-ray to the pattern(s), which has/have been pre-calculated and provided to her. Π has no control over any of the stages during which the pattern(s) are created. She has no other option than to take these patterns for granted. What is important, is that she is also incapable of deviating from the settings of the calibrating filter, which will determine which elements of X-rays will be considered more or less similar to various categories within the reference database. So, if the connection that Π establishes between the X-ray and a certain label will turn out to be not strong enough, then the problem is either in the reference database or the calibration filter she has been provided with. In neither case, the agent Π can be considered as the owner of the mistake [5].

If again, Π were to do K's work, then she would be given a method to ensure the success of the task at her disposal. This method is called *backpropagation* (more on the term see [13]. This is a technique that K uses to make sure that she calibrates her comparison filter properly. This is how it works. In contrast to Π, who projects pre-determined patterns on a new image, K works other way around: she receives from the labeler Λ a database with sets of labeled images and creates patterns that can guarantee that each image matches its respective label. Since K does not know what these labels mean or is able to read the images, she operates on metalevel. She assumes that all images are labeled correctly and catalogs an index of labels matching images. This allows her to remove the labels and carry out comparison *as if* the labels were unknown; the index, however, will help her to find the correct label and measure her success. She then compares images to each other, meticulously scanning every segment for the color densities, agglomerating segments in various ways and averaging the measurements (much like Π). She will eventually arrive at a set of highly processed and codified set of elements, which can be used as a tool (akin to glasses that filter out certain elements) to reconstruct each original image. This set of elements is the prototype of the comparison filter. At first, she will make a guess about the importance of each element for the filter, rendering their "weight"—i.e. the degree of relevance—more or less randomly. Next, K will apply her tool to each image, and, with the help of the index, compare the reconstructed representations with original images. It is important for her to find out how far off the target she got to calibrate the comparison filter. The calibration consists in adjusting the weights of the elements in the filter, little by little, based on the estimation of how far off the correct label the filter has brough her. She is working with a type of a feedback-loop: she compares the filtered result with the target value of 1 (this image is correctly correlated with the label), estimates the discrepancies, and adjusts (recalculates) the weights in the comparison filter. This helps K create patterns to match labels.

If agent Π were able to do all this, it would seem justifiable to ascribe her failure in the sense of (b), i.e. the failure to select a combination of elements in an observation (such as T's X-ray image) that would connect it to a pre-established pattern(s), but *only if* the failure of her filter was not due to the reasons independent of her control (such as an incorrectly

[25] See, e.g., https://deepart.io.

assigned labels; the reference database not being representative of T’s cases).[26] Please note that I do not argue that we *are* justified to do so, but only that we *would be if* relevant conditions were met. This is just for the sake of argument: one still has to prove that the AI-system can be considered as something that is more than a tool in relevant aspects.

We are left with last alternative (c)—described above as problem solving. Given the nature of the limitations on π-ing, it is fairly apparent that the option (c) cannot be properly applied to it. But let’s take a closer look. The patient’s question “What is wrong with me?” is—at least initially and even more so in complex cases—open-ended; it requires exploration and creative search for hypotheses, as well as prioritizing and limiting them. Each hypothesis opens up an alternative pathway of testing, contextualization, and explanation; each requires recourses and time which is crucial especially in emergency situations (see more Stanley & Nyrup [52]. Π has no means for any of these tasks; creative hypothesis search is not available to her as she does not do anything more than comparing pixels in a predetermined context. So, it would not make sense to say that Π failed to solve T’s problem when her judgment about T's condition turned to be wrong. This is simply because she has never tried to answer T’s question.[27] Solving the problem, in this case, would presuppose addressing things that matter to the patient. What matters to the patient, arguably, is that she gets an explanation of her symptoms such that would either clarify her experiences (making it easier to cope, i.e. psychological help) and/or would lead to cure, alleviating or management of her of symptoms (physical help). Searching for such an explanation is a task that goes outside the narrow scope of finding similarities with a pre-selected set of cases [24].

11 The world of Ω: what would it mean to succeed in ω-ing?

To remind, ω-ing in its alternative world reflects the idealized diagnostic practices described in research literature. In the world of Ω, what it means to fail also depends on what it means to succeed. And since diagnosis is a teleological process, its goals define the criteria of success. I will explain this now.

In this context, I will distinguish between diagnosis in the broader sense, as a procedure of determining the causes of patient’s discomfort, and diagnosis in the narrower sense as the end result of this procedure which is expressed in the judgment about the patient’s condition.[28] In what follows, I will be concerned primarily with the broader sense of the diagnosis as a procedure.[29] I will postulate the following general account of medical diagnosis as a complex structured teleological procedure:

> Diagnosis is a complex procedure that is regulated by the rational norms of knowledge contextualization, and consists in the explanation of the symptom(s) experienced by the patient, its(their) underlying cause(s), with the goal of eliminating this cause(s) and/or rendering the condition understandable to the patient in such a way that aids her ability to cope and live with the condition (e.g., through life changes).

To say that diagnosis is a complex procedure is to acknowledge that determining what is wrong with someone’s body or mind is not a matter of exoteric knowledge [47]. There is no crystal ball that can offer the right answer.[30] It is a hermeneutic puzzle (cf. [37], often with multiple

[26] To avoid overcomplicating the discussion, I won’t introduce an additional parameter in the analogy: namely, the role of external constraints under which Π is carrying out her calculations, and the possibility for her to decide on those. These relate to the hyper-parameters of the model, such as the number of layers in the model, types of these layers, type of the activation function, ways to correct bias in the model and to avoid overfitting, type of backpropagation method and its performance metrics etc. This is in itself an interesting thought experiment: How would the ability of the artificial learning agent (in our case DNN) to set its own hyperparameters affect its ability to own mistakes? If the AI does not have this capability, what is the designer's contribution to the mistake, if any?.

[27] One could object at this point by saying that this limitation is easy to overcome: all we need to do is to give Π access to something like an extensive medical encyclopedia. This would allow her to supplement her comparison of T’s X-ray and patterns’ reference dataset with other relevant information. This way she would be able to discard those explanations of T's condition which are less likely and prioritize those that are more likely. In principle, it would be interesting to explore this scenario, however, doing so would make Π analogous to a combination of different AI models rather than only to DNN, significantly complicating the analogy and shifting its focus. One would have to, among other things, explore the details of the new AI model(s) combining DNN’s output with additional clinical data as well as the limitations of this(these) specific model(s). Furthermore, one has to determine how various elements within: which elements are equivalent to her cognitive capacities, and which refer to collective decision making (involving other human agents). Since this paper focuses merely on DNN as a type of diagnostic AI, I will limit myself with the conclusion that one should not expect from this type of models more than they are designed to provide.

[28] For diagnosis as a classification tool, see Jutel [34].

[29] Despite the central role of the diagnosis in medical practice, so far there is no extensive discussion in philosophy about its definition. I base my account primarily on Kennedy [35], Jutel [34], Walker et al. (1990), and Black [9].

[30] Campolo and Crawford [12] offer an interesting angle on de-mystifying AI and its powers.

unknowns, requiring decision under uncertainty [21] and ability to generate and manage the horizon on hypothetical explanations quickly and efficiently Stanley & Nyrup [52]. It is not one action, but includes a number of actions, jointly leading to the same goal. Diagnosis is a teleological procedure, oriented toward treatment. The aim of the diagnosis is an explanation of *patient's complains/symptoms*[31] in terms of underlying cause(s).[32] This is essential because the purpose of the diagnosis—when possible—is to help the patient by curing the disease (i.e. addressing the underlying cause) and/or rendering the condition understandable to the patient in such a way that aids her ability to cope and live with the condition. Kennedy (35, 9) sums this up:

> "Many clinicians consider a good (or 'golden standard') diagnosis to be one that proposes a causal explanation, for two important reasons. First, knowing the cause of a patient's condition facilitates treatment (by allowing for the possibility of intervening on the cause); second, patients tend to both desire and respond positively to explanatory diagnosis rather than diagnostic labels that are not explanatory".

The actions that together lead to a diagnosis as a judgment about the condition of the patient are subject to rational normativity of contextualization of knowledge. This is required to *correctly correspond* the experiences of the patient with the existing body of knowledge about the behavior and properties of various diseases. The rational normativity that guides such complex tasks (consisting of different epistemological tasks) has been translated into a structured procedure which ensures that this contextualization is done properly, and to minimize the risk of mistake due to misunderstanding of the diagnostic purpose. This makes diagnosis an institutionalized procedure, i.e. a procedure subjected to the recognized rules and norms of the medical profession (cf. [34], 64).

So,

> for the agent Ω to succeed in ω-ing, would mean to provide an explanation of the symptoms experienced by patient such that this explanation would make the symptoms understandable in the light of the underlying cause(s) and lead to treatment plan that will either eliminating this cause(s) and/or rendering the condition understandable to the patient in such a way that aids her ability to cope and live with the condition.

12 ω-ing and decision-making

A common approach to the explanation of the diagnosis is via the description of the procedures that are required to properly access the condition of the patient. For example, D.A.K. Black defines diagnosis as "a process of logical construction from the history, examination, and planned investigations" (1968, 51). Three core steps of any diagnosis are: (a) taking the history, (b) performing the physical examination and, when needed, (c) carrying out tests. This account of diagnosis can be considered structural because it aims to describe the proper procedure to aid decision making concerning treatment. The requirement of a certain structure is motivated by—what I will call here—the norms that govern the contextualization of knowledge.

Based on this understanding of diagnostic procedure, I offer the following revised account:

> Diagnostic process is an explanation of the symptom(s) experienced by the patient, whenever possible, through its(their) underlying cause(s) on the basis of patient's complaints (subjectively reported), history (anamnesis), parameters reflecting the objective condition of the patient (such as visual & physical examination, lab tests etc), and the recognized description of the known diseases (incl. their etiology and clinical description), such that this explanation serves the goal to improve the patient's condition (e.g., cure, alleviate or help managing symptoms).

So, to be successful, the agent Ω will have to follow these three steps to concretize the diagnostic judgment in the context of the specific patient, T. First, she will have to (a) take *the history of the illness.*[33] It consists of *the description of the episode of illness* (symptoms) in the context of the background *information about the patient* (initial hypothesis as to the cause of the symptoms). The description of the episode of illness is the overview of the outbreak of symptoms. This must include the report of "the presenting symptom" [9], 51, the reflection of the patient's complain, i.e. the description of what is wrong as it is experienced by the patient him/herself

[31] It must be noted, however, that there are exceptions. Not all diagnoses explain symptoms experienced by the patient. Sometimes, the diagnosis is a result of an incidental discovery or observations which are noticeable for a specialist but not for the patient.

[32] Please note, this is from an ideal point of view, that is to say, it is about what diagnosis should be. In reality, not all diagnoses are causal. This may happen for a number of reasons. For example, in some cases the cause is not known, and the diagnosis may merely be reflective of a (collection of) typical symptom(s), and, again, in some cases the causal interplay maybe be too complex to untangle. This, however, this does not mean that the same standard does not apply to such cases; it merely points to the limitations of our knowledge. That is to say that, should we have the means to identify a cause(s) for such conditions, we would be under the requirement to do so.

[33] van Baalen and Boon [57] refer to this as the "patient's story".

Table 2 A schematic representation of ω-style decision process about the patient's condition

Contextualization in the physical reality of the patient	1. History of the present illness Physical examination of the patient	
Contextualization in the body of knowledge about known diseases	2. Hypothesis about the patient's condition	
	2.a. Testing the hypothesis via laboratory tests (hypothetical diagnosis)	2.b. Judgement about the patient's condition (immediate diagnosis) (This judgment will further be used for the purposes of treatment)
Contextualization in the physical reality of the patient	3. Interpretation of the tests' result in the context of the physical reality of the patient Judgement about the patient's condition (maybe supplemented with another expert opinion)	

(the subjective side), the associated symptoms (the symptoms that may not be of immediate concern of the patient but are normally accompanying the presenting symptoms in different types of conditions), and the system review (an inquiry about the functioning of the other major systems of the body, not apparently linked to the presenting system). "The need for this arises from the misfortune that the presenting symptom may see to point clearly to one system of the body, whereas the seeds of disease are really to be found in another" [9], 33). The symptoms picture should be placed in the context of the patient's history to help form the initial hypothesis about the possible cause of the symptom/symptoms. Among relevant factors are: previous medical history, history of family illnesses, occupational risks, knowledge of other relevant risk factors (such as habits and addictions). The importance of the careful interview piecing together the broader picture of the context in which the symptoms occurred and developed is considered by many the main step toward diagnosis: "about two-thirds of diagnoses can be made on the basis of the history alone has retained its validity despite of the technological advances of the modern hospital" (Walker et al. 1990).

Next, the agent Ω will perform (b) physical examination of T. The physical examination is "the process of evaluating objective anatomic findings through the use of observation, palpation, percussion, and auscultation" (Walker et al. 1990).[34] This allows contextualization in the physical reality of T's body/organism. At this point, the agent Ω will have to evaluate whether the information he has collected is sufficient context for the immediate diagnosis (when cases are relatively straightforward, and the clinical picture is typical). In other cases, she will proceed with a hypothetical diagnosis, i.e. a hypothesis[35] about the underlying cause of the symptoms (called "probabilistic diagnosis" in [35]. She will then put this hypothesis to test on the step (c)—also called the investigation step—through laboratory examinations and—what is of crucial importance—their interpretation in the context of the patient:

> "Accuracy alone is not enough to determine clinical effectiveness [of the diagnostic tests] because the information gained from diagnostic testing does not have a direct effect on patient outcomes—only diagnostic, treatment, and preventive decisions made subsequent to obtaining tests results are considered to have this kind of impact. And these decisions can only be made once the test has been given an interpretation that is medically relevant to the patient in question" [35], 32).

Once a hypothesis is formed, it will be crucial for the Ω to check her assumptions against the reality of T symptoms. Dialog and co-exploration remain the corner stone of Ω's approach: ω-ing starts and ends with T's complains

[34] Palpation is examination by touch; percussion involves tapping a part of the body; and auscultation is listening to a part of a body.

[35] On the role and ways of generating hypothesis in diagnosis see [51].

and experiences: "It allows 'joint authorship' by doctor and patient of the explanatory framework, the therapeutic decision, and the clinical outcomes" [34], 66, cf. also [14, 35], Walker et al. 1990).

The three steps of diagnostic process mark the milestones in the decision making process about the patient's condition, which, when simplified, could look something like Table 2.

13 What would it mean for the agent Ω to fail in ω-ing?

In the world of Ω, we may be justified to say that diagnosis is wrong or incomplete, when it does not explain causally the patient's condition and symptoms and when the treatment, which is proper to the diagnosis (if any is possible), does not lead to a positive change in the patient's condition/experience of the condition. Put differently, Ω's failure to ω correctly is a matter of mis- or not identifying the cause of patient's symptom(s) and subsequent hindrance to the improvement of the patient's condition. This is a more elaborate way to express the basic commonsense idea: *to fail to diagnose is to fail to correctly identify what is wrong with the patient by explaining her symptoms*.

Ω could fail to ω in a number of ways. And sometimes this may happen due no one's fault, just because the cause is genuinely unknown or because he is dealing with previously unknown disease. In the light our inquiry—in connection with the issue of responsibility for a mistake in diagnosis—what is important is Ω's failure due to a mistake. This could happen due to Ω's incompetence, i.e. the lack knowledge or skill which is normally expected from the professional occupying her role. However, we can leave this aside as these cases are relatively straightforward and not very informative if we are to understand responsibility for mistake in diagnosis by/with the aid of AI-systems.[36]

In the context of the structural account of the diagnostic procedure described above, *a failure to diagnose correctly* could be seen as the failure to properly contextualize medical knowledge in application to the specific case of the patients complains/symptoms. This means that in the process of decision making, the agent Ω has failed to take into consideration and/or failed to interpret the relevant symptoms and measurements from the patient, in the light of the knowledge about existing diseases. So, to fail to diagnose (correctly) is to fail to provide an explanation of the patient's complains (i.e. symptoms) in a way that reveals the underlying cause(s) of these symptoms and, when possible, leads to the elimination of the causes and/or to the improvement of the quality of life, if the condition is untreatable. Since such explanation is causal in nature, this, in effect, entails failure to identify the underlying cause (or condition when the cause is generally unknown) in such a way that it would explain and help to treat patient's symptoms.

14 What will be lost if π-ing is identified with the diagnosis?

From T's perspective, the advantages of ω-ing over π-ing as a diagnostic strategy are not hard to see. One important element that Π is missing, compared to Ω, is the reality of T's symptoms. This means that Π's judgment about T's condition will not be contextualized in the reality of T's organism and the experience of the disease. So, as far as T is concerned, Π's diagnosis is will not necessarily be explanatory of T's experiences.

Another element that Π is missing is orientation toward T's health goals. The problem of discrepancy between the diagnostician goals and the patient's own health goals has been well described by Kennedy [35],[37] and the possibility of such discrepancy is quite high in the case of Π/T relationship. Here is one way it can manifest itself: Π's task appears to be exclusionist by design, that is to say, she aims to exclude the possibility of T having certain conditions. The selection of the conditions that T's X-rays is compared to could be based on the consideration of magnitude of danger to T's health. But: Who is to decide what counts as a greater danger and by what procedure (with what reasoning in mind)? For example, if the only thing that matters for the diagnostician is an immediate life threat, is this a satisfactory limitation of health care expectations of a patient? Think about this: T may have different goals, such as preventing chronic conditions, preventing illnesses that stay dormant for years and manifest themselves later in life, or raising quality of life with a chronic condition that, e.g., cause itch or mood swings. Is it justifiable to exclude or bracket out such considerations out of the possible realm of expectations that patients may have from heath care? Reducing diagnosis to π-ing comes with a high risk of, so to say, "losing the patient": making the entire procedure irrelevant to her experience of the disease, her health goals, and, intimately, to the entire spectrum of well-being considerations.[38]

[36] To make sure, I argue that we are not justified to attribute moral responsibility for a mistake in diagnosis to an AI system. However, a way to further this analysis would be to discuss how the dependency of a medical professional on the AI system—in the case of a hybrid agency—affects her moral responsibility for misdiagnosis and new types of vulnerabilities the human agent acquires due to this dependency.

[37] On different conceptualization of models of patient-doctor relationship see, e.g., Emanuel and Emanuel [23]. On participatory medicine and active patient paradigm see e.g., Hood and Auffray [32], Lejbkowicz et al. [38], Fraenkel and McGraw [25], Arnetz et al. [1].

[38] More on the role of AI in the patient-centered healthcare see, e.g., Bjerring and Busch [8], Kennedy [35], Babushkina [3], Dalton-Brown [18], Binns et al. [7].

Moreover, identifying π-ing with the diagnosis, could exclude a multitude of ways to minimize the risk of misidentifying the patient's condition. ω-ing allows for a wide range of cognitive tools to be used in this case, such as understanding (vs mere believing that something is the case) and interpretation (which allows for determining which bits of information are irrelevant for the problem at hand, or under what circumstances they can be considered relevant). Moreover, ω-ing is open to the hypothetico-deductive approach, which allows the preliminary diagnosis to be constantly tested against the patient's physical reality as well as the available body of medical knowledge.In her toolkit, Ω also has the possibility of discussing the assumed diagnosis with other medical professionals to evaluate if any important considerations have been omitted or misinterpreted. On the other hand, Π is blind to her methodology, having no means to evaluate any part of the examination she is performing. This means that she has no way to know the limitations of her methods and materials, margins of error or any means to interpret the results of her calculations. How representative is the pattern of T's case? Do the cases, on the basis of which the patterns were calculated, include cases like T?

As a result, Π's approach runs the risk to significantly limit the range of conditions that will be identified and treated. This could result in (default) preselection of.

(a) convenient (i.e. those that fit the tools),
(b) straightforward cases (easy to identify with the given tools) that
(c) fit certain (silent) criteria of importance (with the possibility of external parties influencing the preselection of conditions for patient's to be compared to).

As a result, for Π there exists a risk of filtering out complex cases, and leaving them undiagnosed, because they do not fit the patterns which were pre-selected for the specific database (excellent examples of this are provided by Kennedy [35] and Jutel [34].

In the end of the day, what T can learn from Π is how similar the state of his lungs is to a number of pre-selected conditions, and only to them. To aid an accurate understanding of how to make decisions concerning T's health based on Π's judgment, Π must inform T that there is a chance that his case may not be typical for the reference database that was used, that she himself does not know how accurate the labels are, and that her judgment only applies to the state of the lungs, and thus cannot be used to explain any other symptom that T is experiencing. Furthermore, failing to inform T about these clauses could hinder the epistemic conditions of informed consent and thus underline her autonomy and ability to judge the relevance of the diagnosis to her condition.

Having said that, I would like to make a few things clear. First, I do not imply that Ω's approach is flawless. But at least her approach allows for dialog, re-evaluation of assumptions, methods used, and suggested causal links, and the possibility of doubt, reconsideration of and contesting[39] the judgments concerning T's condition. My point is that giving up on the cognitive tools available to her should not come lightly as we may risk losing the goal of the diagnostic process altogether. Second, this paper does not argue for the exclusion (or inclusion) of AI systems from the diagnostic processes: this is a subject of a separate inquiry.[40] The goal here is only to contribute to the discussion of their proper use and responsible employment. Various pattern recognition techniques may be invaluable for certain parts of this process, when they are used with full comprehension of their role, possibilities and limitations. But completely substituting the rich diagnostic process with pattern recognition has very high stakes—so we need to be careful not to rush into changes that would potentially have well-being costs for the patients.

15 Where to go from here?

It is important to keep raising the awareness of the personnel using AI-system for the diagnostic purposes about the true capabilities and limitations of such systems, as well as the proper ways of incorporating the output of such systems in diagnostician's decision making process. This is needed if we are to avoid normalizing the lack of responsibility for the possible mistakes in the diagnosis, and we are to prevent the feeling of alienation from the decision process by the human agents.

Besides the unjustified transfer of the ownership of mistake to the AI-systems themselves, there is yet another escape route from moral responsibility—treating misdiagnosis as an accident. And the feeling of being alienated from the decision making process, the loss of control over it, is an easy excuse for going down this way. Easy, but not unjustified, and this is why. For the category of accidents to make sense it should only be used for undesired (harmful) events that were no one's fault, and it would be unreasonable to expect that they are foreseen and prevented. This, *ex vi termini* excludes the situation where the harm has occurred due to such attitudes (and actions) as concealing, ignoring, or unwillingness to obtain or use available knowledge that could help preventing the harmful event. When it comes to AI-systems, this concerns the knowledge about its nature, types of processes it involves, as well its limitations and interpretability of its results. Such knowledge is available, albeit often not adequately presented to the users utilizing AI-systems to achieve their professional goals in medical

[39] On the impotence of contesting AI diagnosis in patient-centric paradigm see Ploug and Holm [42].

[40] For example, Tigard [54] presents arguments against deployment of such high-risk systems in health care.

sphere.[41] So, omitting to use this knowledge *ex vi termini* cannot produce an accident [58].

So, from the ethical point of view, we need to seriously consider *incorporating certain epistemic responsibilities*[42] *in the concept of responsible professional stance* toward such potentially disruptive and high-risk technology as AI.[43] What this would entail is this:

> When the judgment about diagnosis results from AI/medical professional interaction and is primarily motivated by the AI output, all precautions should be taken to exclude the possibility of a mistake in the diagnosis due to the misinterpretation of the nature of the AI system, interpretation of its output, and general over-reliance on it.

In cases when this imperative is violated, we have to further discuss two options. On the one hand, considering the failure to use the available knowledge about the AI diagnostic system and its limitations a type of negligence, when it comes to the medical professionals using AI-based diagnostic systems in their work. For example, Sand et all (2021) have proposed a set of overlapping domains of competency for medical professionals which include: an obligation to understand and critically assess whether AI outputs are reasonable in the context of the specific diagnostic procedure, to know and understand the input data, awareness of ones' own professional skills declining due to the reliance on the AI system, to understand AI's task specificity, an obligation of assessing, monitoring, and reporting output development over time and well as the obligation to inform the patient about uncertainty (sensitivity/specificity rates) involved. On the other hand, there is good reason to start treating the failure to estimate and/or communicate to the professional users the working principles of the system, its limitations, and principles upon which the results must be interpreted for the diagnostic purposes, as a breach of responsible design principles.

Author contributions I am the sole author of this paper.

Funding Not applicable.

Declarations

Conflict of interest The authors report no conflicts of interest.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent to publish Not applicable.

References

1. Arnetz, J.E., Almin, I., Bergström, K., Franzén, Y., Nilsson, H.: Active patient involvement in the establishment of physical therapy goals: Effects on treatment outcome and quality of care. Adv. Physiother. **6**(2), 50–69 (2004). https://doi.org/10.1080/14038190310017147
2. Austin, J. L.: A plea for excuses: The presidential address. In Proceedings of the Aristotelian Society, vol. 57, pp. 1–30. Aristotelian Society, Wiley (1956)
3. Babushkina, D.: Towards Patient-Oriented Transparency. In: Koskinen, J., Rantanen, M., Tuikka, A-M. & Knaapi-Junnila, S. (eds.). Proceedings of the conference on technology ethics, pp. 117–124. CEUR Workshop Proceedings (2020)
4. Babushkina, D.: Robots to Blame? In Nørskov, M., Seibt, J., Quick, O.S. (eds). Culturally sustainable social robotics: Proceedings of robophilosophy conference 2020, pp. 305–315. Amserdam. IOS PRESS (2021)
5. Babushkina, D., Votsis, A.: Epistemo-ethical constraints on AI-human decision making for diagnostic purposes (coauthored with A.Votsis). In: Ethics and Information Technology, 24. The ethics and epistemology of explanatory AI inmedicine and healthcare (Special issue). (2022). https://doi.org/10.1007/s10676-022-0962
6. Behdadi, D., Munthe, C.: A normative approach to artificial moral agency. Mind. Mach. **30**, 195–218 (2020)
7. Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., Shadbolt, N.: 'It's reducing a human being to a percentage' perceptions of justice in algorithmic decisions. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. pp. 1–14 (2018)
8. Bjerring, J.C., Busch, J.: Artificial ontelligence and patient-centered decision-making. Philos. Technol. **34**, 349–371 (2021). https://doi.org/10.1007/s13347-019-00391-6
9. Black, D.R.K.: The logic of medicine. Oliver and Boyd LTD. Edinburgh and London (1968)
10. Blacklaws, C.: Algorithms: transparency and accountability. Philosoph. Transact. Royal Soc. A Math. Phys. Eng. Sci. (2018). https://doi.org/10.1098/rsta.2017.0351
11. Burr, C., Taddeo, M., Floridi, L.: The ethics of digital well-being: A thematic review. Sci. Eng. Ethics 1–31 (2020)
12. Campolo, A., Crawford, K.: Enchanted determinism: Power without responsibility in artificial intelligence. Engag Sci Technol Soc 6, 1–19. (2020). https://doi.org/10.17351/ests2020.277
13. Chollet, F.: Deep learning with python. Manning (2018)

[41] In this respect, Park et al. (2019) may be interesting as they discuss the range of knowledge about AI that medical students should have.

[42] On general account of epistemic responsibility see Code [15], Robichaud and Wieland [45], Corlett [16], more specifically on epistemic responsibility in application to medical profession see, e.g., van Baalen and Boon [57].

[43] On high-risk AI and it regulations see European Commission (2021).

14. Clark, J.A., Mishler, E.G.: Attending to patients' stories: Reframing the clinical task. Sociol. Health Illn. **14**(3), 344–372 (1992)
15. Code, L.: Epistemic Responsibility. SUNY Press, Albany (2020)
16. Corlett, J.A.: Epistemic responsibility. Int. J. Philos. Stud. **16**(2), 179–200 (2008). https://doi.org/10.1080/09672550802008625
17. Cornock, M.: Legal definitions of responsibility, accountability and liability. Nurs. Child. Young People **23**(3), 25–26 (2011)
18. Dalton-Brown, S.: The ethics of medical AI and the physician-patient relationship. Camb Q Healthc Ethics. **29**(1), 115–121 (2020). https://doi.org/10.1017/S0963180119000847
19. de Sio, F.S., Mecacci, G.: Four responsibility gaps with artificial intelligence: Why they matter and how to address them. Philos. Technol. **34**, 1057–1084 (2021). https://doi.org/10.1007/s13347-021-00450-x
20. Dignum, V.: Responsible artificial intelligence: How to develop and use AI in a responsible way. Springer Nature (2019)
21. Djulbegovic, B., Hozo, I. Greenland, S.: Uncertainty in clinical medicine. In: Gifford, F (Ed.), Philosophy of Medicine pp. 299–356. Oxford, UK North Holland (2011)
22. Durán, J.M., Jongsma, K.R.: Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. J. Med. Ethics **47**(5), 329–335 (2021)
23. Emanuel, E., Emanuel, L.: Four models of the physician-patient relationship. JAMA **267**(16), 2221–2226 (1992)
24. European Commission. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. Com/2021/206 final. (2021). Retrieved 16 December, 2021, from https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206
25. Fraenkel, L., McGraw, S.: What are the essential elements to enable patient participation in medical decision making? J. Gen. Intern. Med. **22**(5), 614–619 (2007)
26. Frankfurt, H.G.: On bullshit. Princeton, Princeton University Press (2005)
27. Gerke, S., Minssen, T., Cohen, G.: Ethical and legal challenges of artificial intelligence-driven healthcare. In Artificial intelligence in healthcare. Academic Press (2020)
28. Gogoshin, D.L.: Robot responsibility and moral community. Front. Robot. AI, **8** (2021). https://doi.org/10.3389/frobt.2021.768092
29. Grote, T., Berens, P.: On the ethics of algorithmic decision-making in healthcare. J. Med. Ethics **46**(3), 205–211 (2020)
30. Habli, I., Lawton, T., Porter, Z.: Artificial intelligence in health care: Accountability and safety. Bull. World Health Organ. **98**(4), 251 (2020)
31. Hakli, R., Mäkelä, P.: Moral responsibility of robots and hybrid agents. Monist **102**(2), 259–275 (2019)
32. Hood, L., Auffray, C.: Participatory medicine: A driving force for revolutionizing healthcare. Genome Medicine (2013). https://doi.org/10.1186/gm514
33. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Ng, A.Y.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. Proc AAAI Confer Artif Intellig **33**(1), 590–597 (2019)
34. Jutel, A.: Putting a name to it: Diagnosis in contemporary society. Johns Hopkins University Press, Baltimore (2011)
35. Kennedy, A.G.: Diagnosis: A guide for medical trainees. Oxford University Press, New York (2021)
36. Kudina, O., de Boer, B.: Co-designing diagnosis: Towards a responsible integration of machine learning decision-support systems in medical diagnostics. J. Eval. Clin. Pract. **27**(3), 529–536 (2021)
37. Leder, D.: Clinical interpretation: The hermeneutics of medicine. Theoret. Med. **11**(1), 9–24 (1990)
38. Lejbkowicz, I., Caspi, O., Miller, A.: Participatory medicine and patient empowerment towards personalized healthcare in multiple sclerosis. Expert Rev. Neurother. **12**(3), 343–352 (2012)
39. Micocci, M., Borsci, S., Thakerar, V., Walne, S., Manshadi, Y., Edridge, F., Mullarkey, D., Buckle, P., Hanna, G.B.: Attitudes towards trusting artificial intelligence insights and factors to prevent the passive adherence of GPs: A pilot study. J. Clin. Med. **10**(14), 3101 (2021). https://doi.org/10.3390/jcm10143101
40. Morley, J., Machado, C., Burr, C., Cowls, J., Taddeo, M., Floridi, L.: The debate on the ethics of AI in health care: A reconstruction and critical review (2017). Available at SSRN: https://ssrn.com/abstract=3486518 or http://dx.doi.org/https://doi.org/10.2139/ssrn.3486518
41. Neri, E., Coppola, F., Miele, V., Bibbolino, C., Grassi, R.: Artificial intelligence: Who is responsible for the diagnosis? Radiol med **125**, 517–521 (2020). https://doi.org/10.1007/s11547-020-01135-9
42. Ploug, T., Holm, S.: The four dimensions of contestable AI diagnostics-A patient-centric approach to explainable AI. Artif. Intell. Med. **107**, 101901 (2020). https://doi.org/10.1016/j.artmed.2020.101901
43. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ng, A. Y.: CheXnet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint. arXiv:1711.05225 (2017)
44. Redden, J., Dencik, L., Warne, H.: Datafied child welfare services: unpacking politics, economics and power. Policy Studies **41**(5), 507–526 (2020)
45. Robichaud, P., Wieland, W. (eds.): Responsibility: The epistemic condition. Oxford University Press, Oxford (2017)
46. Sand, M., Durán, J.M., Jongsma, K.R.: Responsibility beyond design: Physicians' requirements for ethical medical AI. Bioethics (2021). https://doi.org/10.1111/bioe.12887
47. Schneeberger, D., Stöger, K., & Holzinger, A.: The European legal framework for medical AI. In International Cross-Domain Conference for Machine Learning and Knowledge Extraction. Springer, Cham (2020)
48. Schönberger, D.: Artificial intelligence in healthcare: A critical analysis of the legal and ethical implications. International Journal of Law and Information Technology **27**(2), 171–203 (2019)
49. Snapper, J.W.: Responsibility for computer-based decisions. In: Goodman, K.W. (ed.) Ethics, computing, and medicine: Informatics and the transformation of health care, pp. 43–56. Cambridge University Press, Cambridge (1998)
50. Stahl, B.C., Coeckelbergh, M.: Ethics of healthcare robotics: Towards responsible research and innovation. Robot. Auton. Syst. **86**, 152–161 (2016)
51. Stanley, D.E., Campos, D.G.: The logic of medical diagnosis. Perspect. Biol. Med. **56**(2), 300–315 (2013). https://doi.org/10.1353/pbm.2013.0019
52. Stanley, D.E., Nyrup, R.: Strategies in abduction: Generating and selecting diagnostic hypotheses. Journal of Medicine and Philosopy **45**(2), 159–178 (2020). https://doi.org/10.1093/jmp/jhz041
53. Tigard, D.W.: Taking the blame: Appropriate responses to medical error. J Med Ethics. **45**(2), 101–105 (2019). https://doi.org/10.1136/medethics-2017-104687
54. Tigard, D. W.: Big Data and the threat to moral responsibility in healthcare. In: Datenreiche Medizin und das Problem der Einwilligung pp. 11–25. Berlin, Heidelberg. Springer (2021)
55. Tigard, D.W.: There is no techno-responsibility gap. Philos. Technol. **34**, 589–607 (2021). https://doi.org/10.1007/s13347-020-00414-7
56. Trocin, C., Mikalef, P., Papamitsiou, Z., Conboy, K.: Responsible AI for digital health: A synthesis and a research agenda. Inf Syst Front (2021). https://doi.org/10.1007/s10796-021-10146-4

57. van Baalen, S., Boon, M.: An epistemological shift: from evidence-based medicine to epistemological responsibility. J. Eval. Clin. Pract. **21**(3), 433–439 (2015)

58. Walker, H.K., Hall, W.D., Hurst, J.W.: Clinical methods: The history, physical, and laboratory examinations. Boston. Butterworths (1990). Available from: https://www.ncbi.nlm.nih.gov/books/NBK201

59. Zimmerman, M.J.: Another plea for excuses. Am. Philos. Q. **41**(3), 259–266 (2004)