Responsibility, second opinions and peer-disagreement: ethical and epistemological challenges of using AI in clinical diagnostic contexts

Hendrik Kempt, Saskia K Nagel

Applied Ethics Group, RWTH Aachen University, Aachen, Germany

Correspondence to
Mr Hendrik Kempt, Applied Ethics Group, RWTH Aachen University, 52056 Aachen, Germany; hendrik.kempt@humtec.rwth-aachen.de

Received 31 March 2021
Accepted 29 November 2021
Published Online First 14 December 2021

To cite: Kempt H, Nagel SK. *J Med Ethics* 2022;**48**:222–229.

ABSTRACT

In this paper, we first classify different types of second opinions and evaluate the ethical and epistemological implications of providing those in a clinical context. Second, we discuss the issue of how artificial intelligent (AI) could replace the human cognitive labour of providing such second opinion and find that several AI reach the levels of accuracy and efficiency needed to clarify their use an urgent ethical issue. Third, we outline the normative conditions of how AI may be used as second opinion in clinical processes, weighing the benefits of its efficiency against concerns of responsibility attribution. Fourth, we provide a 'rule of disagreement' that fulfils these conditions while retaining some of the benefits of expanding the use of AI-based decision support systems (AI-DSS) in clinical contexts. This is because the rule of disagreement proposes to use AI as much as possible, but retain the ability to use human second opinions to resolve disagreements between AI and physician-in-charge. Fifth, we discuss some counterarguments.

INTRODUCTION

In this paper, we discuss the proposal to use artificial intelligent decision support systems (AI-DSS) as providers of second opinions in medical diagnostics from the epistemological and ethical perspectives of peer-disagreement, and provide a rule avoiding most of the raised challenges to such a proposal.

To improve the understanding of what it means to use AI as a source of second opinions, we first differentiate three different types of second opinion according to their clinical use—patient-initiated ones, physician-initiated ones and case-initiated ones. We reconstruct second opinions as expert-opinions with equal epistemological weight as the initial opinion. Under this view, then, any disagreement between initial and second opinion ought to be understood as peer-disagreement.

We also introduce the different types of AI-DSS,[1] and in so doing make apparent that the ability to provide diagnostic recommendations is not matched with the explanation of how these recommendations are formed. The difference in evidence-processing and lack of explainability renders an AI-DSS largely accurate but unchallengeable.

Conflicts between human initial opinions and AI-second opinions, therefore, may not qualify as peer-disagreements, as its 'reasoning' is not reason-based like an expert's evaluation. Hence, we cannot peer-disagree with an AI-DSS, leaving a responsibility gap when trying to decide what to do in case of a conflict. The physician-in-charge, then, needs to settle on a diagnosis based either on the recommendation of the AI-DSS they are not able to confirm, or reject the AI-DSS-recommendation without being able to sufficiently explain why. This shifts the burden of proof towards the physician-in-charge and raises problems regarding the attribution of moral responsibility.

Given this challenge, we weigh the arguments from accuracy, that is, the fact that AI-DSS will soon be equal or superior to the average physician, against the normative demand that clear attributability of responsibility should be conserved.

In order to use the efficiency and accuracy of AI in diagnostic processes without undermining the clear attribution of responsibility, we propose the 'rule of disagreement'. This rule permits AI as second opinion providers unless they deviate too far from the initial opinions, in which case a second human opinion ought to be sought. If the AI-DSS confirms the initial opinion, however, no further steps need to be taken.

COOPERATION IN DIAGNOSTIC SETTINGS

Cooperation among agents is ubiquitous in modern societies. The main feature of cooperation is the pursuit of a common purpose or goal by sharing physical and cognitive tasks that often could not be achieved through actions by any agent alone (or only at some considerably increased costs). The condition of a common purpose is what sets cooperation apart from mere 'coordination', in which agents communicate their own goals with each other and adjust their actions so everyone can separately realise their respective goals (eg, in traffic).

In the medical field, cooperation has been standard practice ever since the specialisation of the medical profession began requiring a shared labour approach to diagnosis and treatment, and has been increasingly so ever since (cf. Molleman *et al*[2]). The shared goal—correctly diagnosing and properly treating patients—has become such an overly complex process that it can only be achieved by splitting the medical professions into specialisations. To perform a surgery, for example, different specialised medical professions are required to cooperate, from radiologists to anaesthesiologists, nurses and surgeons in the actual operation room. They all share responsibility for the success of the surgery, because their correct cooperative contribution constitutes a necessary condition for the success of the surgery.

In the context of correctly diagnosing a patient's ailment, a certain amount of cognitive labour is shared, akin to the roles of anaesthesiologists and surgeons in surgeries. And akin to the shared responsibility in the cooperative physical labour of surgeries, the shared responsibility in the cooperative cognitive labour of the medical diagnostic process applies to everyone involved, as any physicians' incorrect contribution endangers the correctness of the overall diagnosis and thereby the well-being of the patient. In the following, we are first exploring the form of more or less formalised cognitive cooperation in diagnostic processes and assess how practices of sharing responsibility work.

SHORT TAXONOMY OF SECOND OPINIONS

The main term for institutionalised cooperative actions in the diagnostic process is the 'second opinion'. While we acknowledge that several cooperative diagnoses do not fall under the practice that is commonly associated with the term, we use 'second opinion' as a catch-all term for cooperative decision-making in diagnostic processes (a similar term may be 'four eyes principle'). Two reasons speak for this terminological choice: first, the commonly associated meaning of the term indeed represents many instances of institutionalised cooperative decision making. Second, these institutionalised associations distinguish the cooperative forms of diagnostic teamwork to the mere 'advising' roles of diagnostic exchange that is very common (cf. Mamdani[3]).

From this cooperative perspective, we identify three different types of second opinions that occur in clinical processes:

First, it appears that the most common type of diagnostic opinion as 'second opinion' functions as a patient-initiated quality control (cf. Payne *et al*[4]). In this process, a physician-in-charge proposes a diagnosis to a patient, which is based on the available evidence and the status of the physician as expert of collecting and assessing this evidence. The patient may request a second diagnosis from a different physician to improve their own trust in the diagnosis[5] or determine the adequacy or necessity of a certain treatment plan. A second opinion, then, can either confirm the initial diagnosis based on the available evidence, disagree with the diagnosis based on the evidence, or point toward the insufficient amount of evidence rendering a diagnosis insufficiently supported by evidence to propose as a solution with confidence.

All three of these results of the second opinion towards the initial diagnosis will, theoretically, improve the overall or final diagnosis: If the second opinion is concurring, the fact that two experts conclude the same (or very similar) results with the available evidence will bolster the initial diagnosis as the best available diagnosis. If the second opinion points out a lack of supporting evidence, it can help to improve the initial diagnosis by increased evidence collection and as an epistemological check to ground the diagnosis on sufficient evidence. And if the second opinion disagrees with the initial result favouring an alternative diagnosis, the resulting conflict of the peer-disagreement suggests further rounds of discourse and evidence assessment (see also the findings on how second opinions change the initial diagnosis,[6]).

Second, there is physician-initiated second opinion. This is a request for an opinion to assess the physician-in-charge's proposal, or by relying on the second physician's special expertise for a certain diagnostic issue. Often an informal process based on the physician's confidence in their own diagnosis or diagnostic capacities, the goal of this second opinion is the increase of confidence levels by calibrating the original diagnosis through comparison, and possible incorporation, with the additional one.

While this process is most often of informal nature, we can expect that both the requesting and assisting physician treat this request with the same epistemic diligence as if they were providing the initial opinion. That means the requesting physician is providing all evidence available to ensure the additional diagnosis is as precise as possible (and not holding anything back, to potentially receive the result desired). It also means the assisting physician treats the diagnosis with no less care and attention than if they were providing the initial diagnosis. In practice, this form of second opinion is often hard to distinguish from a mere consultation with another physician, as those informal exchanges, between radiologists, for example, are part of everyday clinical life. It is important, however, to note that any request to provide a diagnosis to a case ought to be treated with the same diligence as if the second opinion was an initial one.

Finally, a third kind of second opinions can be identified. This can be understood as the legal or institutionalised second opinion. In this, we can identify two different versions of legal second opinion. One is the automatically triggered second opinion of an expert to check the diagnosis of the first physician, for example, in cases of breast cancer screenings in Germany (cf. German Federal Ministry of Health[7]). This process is motivated by the relatively high deviance among doctors and the rather impactful consequences of false positives, that is, cases in which a potential source of breast cancer is indicated while it is a false diagnosis (like a biopsy or, possibly, an amputation), or false negatives (ie, a delayed cancer detection and potentially worse course of disease and treatment).[8]

The other version of those institutionalised second opinions occurs in diagnostic centres and 'medical conferences', in which an interdisciplinary group of physicians contributes their perspective and expertise to the best diagnosis for a patient.[9] These are usually implemented for specific diseases, like certain types of cancer, in which the diagnosis and treatment-recommendations ought to be assessed by all relevant disciplines. The same ethical requirements of epistemic diligence apply to any physician involved in these cooperative diagnostic processes as if they were diagnosing them on their own, as failure of any physician involved will potentially have the same severe negative consequences.

One emerging fact from reconstructing these types of second opinion, even with the more informal ones, is that the shared cognitive labour contributes to the confidence levels of the physician-in-charge in providing a certain final diagnosis from which treatment is recommended. Without the input of diagnosis-suggestions of other physicians, the confidence levels would not change, as the physician-in-charge is left with their own diagnosis. In all the mentioned forms, from patient-initiated and physician-initiated ones to case-initiated or generally institutionalised processes of cooperative diagnostic efforts, the shared cognitive labour and requirement for epistemic diligence is indicative of a shared responsibility structure.

RESPONSIBILITY AND DISAGREEMENT

For most clinical purposes, the legal responsibility and liability remains with the physician-in-charge, even though the second opinion may affect the final diagnosis in one way or another. The reason for such a concentrated role-responsibility with the physician-in-charge may lie in the practicality to create legal liability while still encouraging cooperative diagnostic procedures to take place (see Carrick[10]). The role of the physician-in-charge creates this responsibility, as the physician is not only the person to propose the initial diagnosis, but also the one to settle on the final one from which treatment is recommended.

Thus, many philosophical considerations have no immediate practical consequences as the long-standing legal setup is covering some of the concerns presented here (including the general uncertainty of diagnoses). However, this does not mean that these concerns ought not to be taken seriously, as they may inform changes to the legal setup down the line. Our main question of moral responsibility emerges in cases of disagreement among the initial and the second opinion. Without a disagreement, the physician-in-charge has no reason to assume they could be mistaken, as all available evidence and the physician's own diagnosis are reaffirmed by the second opinion, independent of the correctness of said diagnosis. As far as responsibility goes, physicians are epistemically justified in their diagnosis if another physician comes to the same conclusions, barring unusual circumstance. A disagreement between initial and second opinion, however, establishes the burden of proof as falling on the physician-in-charge: as the bearer of responsibility for the final decision, their disagreement with a peer-opinion on the same diagnosis ought to be justified.[11 12] [i]

Independent of whether a second opinion is correct or not, the physician-in-charge may not dismiss it without justification. As we have previously assumed that any diagnostic suggestion ought to be made with the diligence of an initial diagnosis, we can categorise these disagreements as peer-disagreements[13]: epistemic equals differ on a question of how to interpret evidence and reach conclusions based on that evidence.[14] In the discourse of social epistemology, this position is called the 'equal views principle'.[15] This position is required to reconstruct peer-disagreements, as the views of two epistemically similar agents, based on the same evidence, ought to be considered equal in weight and thereby diminish the confidence one has in their own opinion if one's peer disagrees. As doctors providing diagnoses are usually understood to be experts with equal standing, the equal-weight-view explains why disagreements between doctors are peer-disagreements.

While this view is agnostic on the conditions of who counts as an expert in general, the specific circumstances of clinical diagnostics are *not* agnostic. Only specifically educated and licensed medical personnel (ie, physicians with full license to practice) may count as epistemic equals in diagnostic contexts, and diagnostics, as pointed out above, has been determined as a process of providing explanatory reasons for diagnoses.

Therefore, a key feature of those diagnostic equal-view peer-disagreements is that they are based on reasons, and their resolution on an exchange of those reasons. We expect medical experts to form their opinion on shareable reasons that can be challenged by their peers—in these cases we can with Weinstein talk of 'epistemic expertise' (pp62–64).[16] The reason why we speak of 'experts' in the first place is not their privileged mental access to evidence, but their skill to solve certain tasks by assessing evidence, forming conclusions based on that evidence, and knowing the parameters that are unknown and how to make them known.[16] They can provide reasons for the weighing of evidence in their conclusions, based on the state of the art research in their respective field and their experience in inferring correct conclusions from available evidence.

A disagreement between the initial and the second opinion can thus be resolved by an exchange of reasons for and against certain diagnostic conclusions. Thereby, the final diagnosis is ideally an amalgamation of the initial diagnosis, a second opinion, and, in case they differ, the epistemic justification of why the final diagnosis and the second opinion differ and why one of them should be favoured.

USING AI TO PROVIDE SECOND OPINIONS

The emergence of semi-automatic diagnostic tools, and the progress made in the development of fully automated AI-diagnostic systems, create an ethical urgency to clarify and justify their potential use. With their improvements potentially surpassing human medical experts in diagnostic accuracy, it appears morally imperative to use their potential in medical diagnostics, even when some questions of responsibility remain open. We acknowledge that most AI-DSS still have considerable hurdles to overcome in order to reach consistent levels of both precision and reliability; yet, for our purposes of discussing disagreements, the need for AI-DSS to outperform physicians is not as relevant, as disagreements can be anticipated nonetheless.

The proposal we are concentrating on is to use diagnostic systems as automated second opinions (eg, Luxton[17]). The core thesis of this proposal is that AI-DSS could replace human doctors in providing additional diagnostic services, akin to a second opinion. This proposal draws some attention as it fully incorporates the potential of AI in the medical diagnostic context, which consists in its fast, reliable and precise diagnoses,[18] its resource-efficient maintenance, and in some cases its utilisation of ever-increasing and improving databases. Additionally, the institution of 'second opinion' as a source of uncertainty-reduction and improved diagnosis is a well-defined starting point for increasingly autonomous AI-DSS to take hold.

Using AI-DSS this way, however, implies the replacement of human cognitive labour, that is, the labour of physicians providing a second opinion, while not completely replacing human cognitive labour and responsibility altogether, as the physician-in-charge remains the decider of the final diagnosis, avoiding the ethical conundrums of outright automating the diagnostic process altogether (see for an elaboration of the lack of 'meaningful human control' in Braun *et al*[19]).

From this, a strong argument can be formed that concludes a moral imperative to use AI in medical diagnostics, which is also guiding the proposal of using AI as second opinion or some other sort of substituting diagnostic input. The anticipated superiority of machines in several diagnoses such as melanoma[20] or early breast cancer detection[21 22] (for an overview of studies on how physicians fare with AI diagnoses, see Topol[1]) suggest that if we rejected their use, we would prefer worse outcomes in diagnostic precision merely because of a lack of justified procedural use of the technology. In fact, some hospitals have moved to replace the routine-controls done by humans with AI, solidifying the concrete urgency of clarifying ethical challenges to diagnostic reality.[23] To answer those issues, we provide rules governing the ethically justified use of AI in replacing human cognitive labour in diagnostic processes.

[i] We thank a reviewer for suggesting to clarify that we only consider a subsection of peer-disagreements: peer-disagreements can occur on most any topic two experts have epistemically justified, different stances on. Our reconstruction of diagnoses as peer-disagreements, however, presupposes disagreements about specific cases with practical consequences, rather than mere theoretical disagreements. Hence, our conclusions are also limited to this understanding of particular peer-disagreements.

Levels of AI-DSS

Before turning to the proposed use of AI in diagnostic systems, we first ought to distinguish the different forms of AI tools and which ones are relevant to the issues of replacing human second opinions with algorithmic ones. We will use the distinction proposed by Topol.[1] His proposal distinguishes diagnostic AI tools in five levels according to their autonomous capacities. As we discussed in the rather informal second type of second opinions, not every feedback from another physician is a second opinion, as we defined second opinions as the provision of a second diagnosis or a criticism of the initial diagnosis based on the evaluation of the available evidence.

The levels that Topol discusses begin with providing basic supplemental information, empowering the physician to make better informed diagnoses. The second level allows for autonomous AI-recommendations of previously identified subproblems that may help form an overall diagnosis of a patient. The third level is called a conditionally automated system in which complete diagnoses are automated but remain under supervision and final decision by a physician. In the fourth level, highly automated systems gather evidence and form diagnoses on their own, while the decisions about treatment remain in human hands, while in level five, the fully automated system takes over the complete diagnostic and treatment process of a patient.

In order to speak of ‘AI as second opinion’, then, we have to assume that this AI is capable of instantiating certain autonomous features that are comparable with the human cognitive labour it is intended to replace. This dismisses mere decision-support systems of Topol’s first and second level, in which, comparable to another physician giving mere advice or additional information, the physician-in-charge is the only source of diagnoses altogether. In turn, however, level three and four of Topol’s order suggest that those are at least capable of forming their own diagnosis if fed with the same evidence that the physician had available, or even more gather its own evidence necessary to form a conclusion.

AI-DSS as cooperation partners?

As the epistemological and ethical conditions of second opinions provided by humans have only partially been researched, it appears even more relevant to describe the changes to that process when introducing AI. We identify two main relevant changes to the clinical processes if AI was implemented as a second opinion: (A) the replacement of human cognitive labour and (B) the emerging human–machine cooperation (from A) between the physician-in-charge and the AI diagnostic system.

In response to (A), replacing human cognitive labour is not generally ethically worrisome. The use of computers has arguably improved our lives by freeing up cognitive capacities we would have otherwise used on calculations. The difference between replacing cognitive labour in the general case with computerised machines and the emerging technology of autonomous AI, however, lies in the latter’s capacity to not only do mechanical calculations faster than any human ever could, but also to imitate the human ability to draw inferences. It is, thereby, capable of simulating an actual replacement of human inferential cognitive labour rather than mere mechanical calculations, with similar unexplainable ‘mental’ decision-making pathways as humans.

The relevant difference between AI and experts in this decision-making process is, as pointed out before, that medical explanations by experts are based on reasons. When medical experts explain their diagnosis, they do not refer to their mental pathways and synopses, but the evidence and theories with which they inferred the best explanation of a patient’s symptoms. We also do not accept an expert’s opinion merely because they think they are correct without telling us why (besides an appeal to second-order evidence, such as track-records, which is a global argument and does not apply to specific disagreement-cases). Similarly, then, an AI’s diagnosis cannot be ‘explained‘ in the same sense by basing the explanation on the vast training data and the algorithm’s isolated, largely unintelligible and—often for patent-protection—clandestine connections in the pattern-detection process. And without reasons, even the successful simulation or imitation of diagnostic behaviour reduces the acceptability of its diagnoses.

If we could fully explain a machine’s workings, this issue would be less worrisome. Explainability is best understood as a matter of degree. With that, several approaches to explaining machine learnt/deep learnt AI can illuminate some of their inner workings. The concept of ‘black box algorithms’, in which decision-making processes are fully unintelligible, thus has become somewhat misleading, as we are in fact able to interpret some decision-making processes of AI and are not fully left in the dark about an AI’s pattern-recognition. From simple strategies to change some input parameters to test for certain outcome-changes[ii] to elaborate strategies like LIME,[24] the lack of full explainability is not to be equated with a lack of knowledge about the system altogether (for an overview of some ongoing debate, see Tjoa and Guan[25]).

However, we have doubts on whether these explainability-efforts will reach levels sufficing as ‘reasoning’ that would allow machines taking a physician’s place in providing second opinions: AI, other than, for example, simple causal decision trees that are not applicable to complex diagnostic situations, do not provide causal explanations as the system is not designed to find causal connections but rather correlations and associations in the training data.[26] The resulting ‘explanations’ rather enable clinicians to interpret the system’s output than to provide complete explanations or causal reasoning of the diagnosis itself.[27 28] Thus, making an AI-DSS more explainable usually merely means making it more interpretable.

This concern can be ameliorated through an increased effort in different interpretative models, as the previously mentioned LIME, in which statistical summaries of key features for an algorithm are presented. These ‘key features’ come relatively close to what we might mean with ‘reasons’.

However, when asking a peer for their reasoning of how they arrived at a particular diagnostic proposition, physicians specifically ask for their weighted explanations of their reasoning. Using interpretable models of an AI, like LIME or similar, however, requires the doctor’s assumptions about the results to play into their interpretation of the machine’s results. An interpretable AI, thus, is not providing *reasons*, but allows for the interpretation of its inner workings *as reasons* for the physician-in-charge. This marks the fundamental difference in which explainability and interpretability of a model (or rather, their results) differ on their ability as counting as part of a peer-disagreement: an interpretation is being done with assumptions from the interpreter, an explanation is made on the grounds of the explanandum. Thus, an interpretable model may be especially useful as additional evidence, but

[ii] We thank an anonymous reviewer for this suggestion.

not as an independently accessible control-institution to the initial opinion (ie, as second opinion). Rudin[29] discusses these issues as (un-)explainable AI is often not able to be combined with contextual information outside the database, which diagnostics usually require.[29]

To B)): First, we can ask whether machines and humans can, in fact, cooperate. As we stated in the beginning, a cooperation is a collective effort towards a shared goal. We can question whether machines have goals, because the answer is not trivial: if they do not have goals, humans and machines simply do not cooperate. But if that is the case, then the nature of an AI 'second opinion' changes from how we have characterised it above: it cannot be a straight substitution, as second opinions are a result of human–human cooperation, and replacing a human with a machine will not lead to the same cooperation, but a less involved human–machine interaction. However, this is merely a conceptual concern of how to properly describe the interactions between initial and second opinion.

Second, depending on the answer to the question of how a less involved human–machine interaction can replace human–human cooperation, the normative question of shared responsibility in these contexts emerges. While the provider of the second opinion ought to be able to be held to the same standards of epistemological diligence, replacing a human with a machine will have effects on how this epistemological diligence translates to shared responsibility: if the machine is not capable of bearing any responsibility, does the responsibility distribute at all? This point appears to have been discussed on a wider basis in the proposals to replace humans with machines in diagnostic settings,[30] as it appears to be an urgent practical issue: how is the relationship between the physician-in-charge and the machine-as-second-opinion to be understood, if before the physician was in a professional relationship with a human colleague as provider of a second opinion on a similar case?

REQUIREMENTS OF PEER-DISAGREEMENTS

The change of burden of proof for the physician-in-charge is significant: a physician having a disagreement with another physician can enter an exchange of reasons, and potentially resolve the issues, or conclude that they require more evidence for a resolution. A peer-disagreement can thus be resolved by convincing the other side (or being convinced by the other side) through the exchange of reasons. A physician having a disagreement with a machine can put forward all the reasons they want, the machine will not react to it on the basis of reason, but merely on the basis of the data. The way the machine interprets the data, however, cannot be changed by the physician-in-charge (eg, through the manipulation of certain toggles or parameters): as this will only lead to a self-confirming bias with the physician changing parameters for data interpretation until the machine gives 'sensible' results to the physician, which already lie within the range of assumptions of the physician in charge. The point of the proposal is, after all, to provide a check on the initial diagnosis, not merely confirm it.

And since the intention is to check the initial diagnosis, the physician ought not to ignore the second opinion as misguided or false; yet they also cannot overturn it in any meaningful way based on an exchange of reasons, as the AI is not only non-explainable as required, but also not receptive for diagnostic reasons. Thus, the burden of proof of the diagnostic process moves from an epistemic equilibrium (in the equal weights view-sense) between the initial and the second opinion, towards the proponent of the initial opinion.

The structure of a physician–physician relationship and its proposed replacement as physician–machine relationship suggest these are similar, the ethical conditions of these structures suggest that they are not. Therefore, based on the requirements of a second opinion as a part of a cognitive cooperation between two potential bearers of responsibility, we can reject the proposal of merely replacing human second opinions with machines.

ACCURACY AND EXPLAINABILITY

One central argument for using AI-DSS without having fully explainable reasoning behind its results (and thereby without peer-disagreement) is its presumed superior accuracy over (the average) physician.[18 31 32]

This argument is supported by two reasons that we ought to briefly discuss here. The first one, brought up in Braun *et al*[19] (pp6–7)), is concerned with the accuracy of human diagnoses. Most diagnoses exhibit a huge deal of uncertainty. Thus, deviating diagnoses are most often not specific disagreements, but differ in weighing and evaluating options. Braun further contends, contra our conclusion in the previous section, that automated diagnostic processes will not shift the burden of proof from the machine towards the physician, as the physician's task of balancing different confidence levels for certain diagnoses and taking responsibility in the decision-making of treatment remains unchanged.[19]

It is, on this view, much less an issue of replaced cognitive labour, as we have reconstructed above, but one of risk assessments in the light of uncertainty for the physician-in-charge, as they are balancing the input to fit their own diagnostic assumptions. This view appears to differ in the conceptualisation of second opinions in general, because the same can be said about human second opinions.

If second opinions are not understood as a collaborative effort but as mere additional evidence, then both a deviating automated as well as a deviating human second opinion are not 'disagreements' in a peer-disagreement sense, but merely contradictory evidence to the initial diagnosis, rendering the issue of peer-disagreement mute. However, as the reconstruction of second opinions with the equal-weight view has shown, both in theory as well as in practice, second opinions are taken to be peer-assessments with the capacity to be compared as equal diagnosis with the initial one. And with the question being whether AI-DSS may function as such peer-assessment, reconstructing these conflicts not as disagreements but as mere contradictory evidence is inappropriate.

In the second argument, contrary to the first point, we may argue that the precision and efficiency of such AI-DSS, presumably on par with many expert skills of physicians, ought to be considered practically equal to a second opinion (see, for example, Kompa *et al*[32]). AI-DSS must pass certain standards of precision to be certified. In this sense, then, a certified AI-DSS and a physician are not different in terms of qualifications, and with precision, speed, and efficiency being paramount reasons for deploying AI-DSS, the argument can be made that the presented normative concerns are overblown[31] or at least outweighed by the moral benefit. The proven accuracy of a machine can, in this regard, count as a reason to reduce the responsibility of a physician as the certification standards can be sufficiently high to support their use. If a machine is wrong and the physician follows its wrong diagnosis, then the physician is somewhat less responsible for the misdiagnosis, as the accuracy of the machine is reason to assume its better 'judgment' of the case.

Any potential responsibility gap emerging from this use of AI, then, is merely a theoretical worry for which we have developed practical rules in similar cases (mainly for pharmaceuticals, in which the specific causes of effect are often only correlated, while tolerating considerable risk of side effects)—the responsibility dissipates into institutionalised rules of compensation or individualised risk-responsibilities as we collectively agreed on a certain amount of uncertainty that we may fall victim to. If the distribution of responsibility is not an issue, it appears unreasonable not to use AI as a second opinion in clinical contexts given their utility.

While proven accuracy is a practical reason for believing that AI-DSS may provide accurate diagnoses, resolving disagreements requires explanations as reasons—otherwise a disagreement is not resolved but merely decided in favour of one side, and AI-DSS may not provide explanations in the required sense. The issue of explainability is widely discussed in the general field of AI ethics,[27 28] because the explanation for a certain behaviour intuitively requires more than probabilistic assumptions about certain connections within the neural network of the AI. Even authors like Grote and Berens[12] accept that explainable AI can be a bigger issue than imprecise AI for the ethical justification of their use in clinical diagnostics (p3). Thus the success-rate of the AI-DSS and subsequent risk-assessments of their use is not the only defining ethical problem, but also the lack of any independent explanation for the specific diagnostic predictions—and in cases of disagreements, as shown above, the limits on explainability can outweigh risk-considerations based on the requirements of attribution of responsibility. Ultimately, AI-DSS do not act, but behave. And precise behaviour is insufficient for attributing responsibility, as only actions, not behaviour, can be explained with reasons. The acceptance that a machine may be statistically more likely to diagnose correctly and claim this to be the case[32] does not absolve the physician-in-charge from making a diagnosis themselves, as otherwise nobody could be made responsible for any mistake in the diagnostic process.

Without explanatory reasons for certain conclusions, a disagreement between a physician and a machine may not be resolved responsibly. It does, however, move the burden of proof towards the physician-in-charge, as they have to incorporate AI-DSS as evidence, but cannot reject the evidence even if they think it is wrong.

RULE OF DISAGREEMENT

The main issues we have worked out so far are that the anticipated use of AI-DSS as second opinion replaces human cognitive labour and, at the same time, requires human–machine cooperation with problematic distributions of responsibility.

We propose a rule for using AI-DSS that would harvest the positive reasons without encountering the question of attributing responsibility in cases of misdiagnoses, and avoids the issue of peer-disagreement in case of conflict:

> **The rule of disagreement:** If a diagnosis provided by an autonomous AI diagnostic system contradicts the initial diagnosis of the physician-in-charge, it shall count as disagreement requiring a second opinion of another physician.

This proposed rule resolves the previously encountered issues. As we have encountered in the case of the legally required second opinion in breast cancer screening, issues only arise if the second opinion is in conflict with the initial one. If the second opinion is concurring or offering the same answer as the initial diagnosis with a sufficiently high confidence level, a discourse around the reasoning of the second opinion is not required. If the machine diagnosis is in conflict with the initial diagnosis, a second human opinion ought to be requested to resolve the issue. This human second opinion functions as conflict-resolution to avoid the problematic attribution of responsibility to machines in the case of misdiagnoses, as not only a machine, but two physicians confirmed the diagnosis.

An analogy to human second opinions might be helpful here: a physician in charge (P1) seeks a second opinion from colleague (P2). P2 provides a disagreeing second diagnosis, but dies due to some accident before they can enter a discourse to resolve the disagreement. Assuming that P1 asked P2 due to P2's track record of precision, we would not allow P1 to take on P2's diagnosis, as P2 cannot be a bearer of responsibility anymore.

In cases of a confirming diagnosis, AI-DSS can thus be treated as confirming evidence, replacing the need for a confirming human second opinion, without having to treat the AI-DSS as a participant in a discourse, as would be required in a case of peer-disagreement. This pragmatic rule of disagreement enables physicians and patients alike to not have to *trust* a machine but expect *reliability*, just as they would with any other tool. At the same time, the autonomy, potential self-learning and improving capacities, and potential responsibility-issues are avoided by keeping physicians in the loop and as exclusive bearers of responsibility.

APPLYING THE RULE OF DISAGREEMENTS

Turning back to the three types of second opinion, we can now assess how the proposed principle of disagreement fits in the described contexts of providing second opinions and why, despite rejecting the role of AI-DSS as full second opinions, they can still vastly improve the precision and quality of the diagnostic process.

First, patient-initiated second opinions are often a sign of distrust in the doctor's diagnosis in a low-stakes environment. Having an AI-DSS review the evidence and confirm the initial diagnosis can provide some trust in the initial diagnosis as well, while avoiding the misconception of AI as a miracle machine (a concern some authors have worried about[33]). Without a conflict between the doctor and the (potentially independently operating) AI-DSS-second opinion service, no further action needs to be taken, while saving resources from physicians having to provide a second opinion to a patient. With a conflict, however, the anticipated benefit of the AI-DSS is lost and a second human physician is required to provide a second opinion.

The use of AI-DSS as the second type of second opinions may work similarly. If a physician approaches an AI-DSS for a second opinion or as an evaluation of their initial opinion, the confirmation can be considered an increase in the confidence levels of the diagnosis, while a conflict ought the physician turn towards another human physician for further evaluation. As this type of second opinion is mainly informal, it does not encounter the issue of peer-disagreement, even though the issues of resolving any conflict between the suggested initial diagnosis and the second opinion remain.

Finally, for the third type our proposal works as well: in the institutionalised version of second opinions triggered by certain cases, using AI-DSS to evaluate the initial diagnosis appears unproblematic with the proposed principle of disagreement. In breast cancer screenings this approach appears beneficial[34] as it does as a check on the recommendations of a diagnostic council, even though it is questionable whether AI-DSS will be an addition to those councils in the foreseeable future. If an AI-DSS is in fact disagreeing with the proposed solution, consulting a physician to resolve the issue is required.

FURTHER CONSIDERATIONS

With any automation of previously high-skilled human tasks comes the risk of de-skilling and over-reliance of the remaining workforce. In the case of de-skilling, this does not seem to be the case in diagnostic contexts if the principle of disagreement is applied properly. The education and experience of radiologists and other physicians remain unaffected as their skills will be required to fulfil the tasks of proposing the initial diagnosis or function as a human second opinion to resolve potential conflicts between an initial diagnosis and an AI-diagnosis.

However, the risk of over-reliance is more imminent. The more autonomous technologies become that imitate human tasks, the more people may be inclined to rely on the technology to take over similar tasks it is not intended to take over. Cases of misuses of the autopilot-function in stage three or four self-driving cars is an example having cost lives in the last few years.[35] In these cases, it was not the technology that failed (as it was not certified to be fully autonomous), but the drivers that mistreated the technology by misusing the technology for tasks it is not certified for. A similar over-reliance may be encountered in providing physicians with a quasi-autonomous AI-DSS that may only be used after the physician has done their own diagnostic work.

The joint statement of the radiological societies in the USA, Canada and Europe, reiterated that any AI-DSS ought to keep physicians in the loop.[36] The professional ethics of physicians demand to be able to take responsibility of the final diagnosis and the accordingly administered treatment. This includes the ability to justify one's own diagnosis which, in turn, would require the physician not to turn to the AI-DSS for an initial diagnosis, as AI-DSS still do not provide reasons for their results. The fact that some physician may use the AI-DSS against the proposed rule of disagreement, however, does not invalidate the rule but rather requires accompanying extensions of the professional ethics.

Finally, keeping physicians in the loop of diagnostic processes means to decrease some of the anticipated advantages of such autonomous technology. The resource-efficiency of fast and precise analysis within a few seconds will not materialise if AI-DSS will only be used in the second instance. Similarly, ameliorating expert-scarcity will not occur if those systems are not to be used as providing an initial (and possibly sole) diagnosis. It appears that if no doctor is present, the progress of this technology remains unavailable, potentially furthering the divide between well-supplied communities and desolate ones.

However, the proposed principle operates under the assumption of the presence of experts and is motivated by problems of responsibility in cases of misdiagnoses and disagreements. If AI-DSS were to be developed that in fact are usable diagnostic tools for laypeople in areas where doctors are not available, the imperative to provide medical care outweighs the concerns for potential harm of misdiagnoses in such medical care. Thereby, the limiting effect of this principle for the general availability of AI-DSS only applies to areas where the careful distribution of responsibility outweighs the necessity for a diagnosis in the first place.

CONCLUSION

Finally, the discussion about finding rules to incorporate AI-DSS as second opinions in the diagnostic process may suffer from a mismatch in the static idea of the role of physicians, and the dynamic idea of AI-DSS. We argued for an implementation of AI-DSS in the diagnostic process that keeps the role of physicians virtually unchanged as the first proponent of a diagnosis that, if challenged by an AI-DSS, will have to be resolved by another physician. However, the development of AI-DSS is dynamic and points towards an increased skillset able to potentially replace physicians in several contexts.

Some authors have started using the term 'partnership' (cf. Patel *et al*[37]) to stress that not only the technology has to fit our established norms of diagnostic decision making, but that those norms also have to incorporate the ever-growing skills of the AI-DSS. This approach aims to both guide the development of AI-DSS to fit the diagnostic process, but also to reimagine the role of physicians to use the AI's full potential in aiding the diagnostic process.

The necessary conversation about changing self-conceptualisations of physicians in the diagnostic process, however, can only be pointed towards here. Such a conversation is well underway within the respective professional associations[38] and may provide a path towards a development of AI-DSS that is not focused on merely replacing human cognitive labour, but to aid the cooperative nature of diagnostics processes even further.

As we have shown here, however, there are ways of providing space for AI-DSS that allows for an incorporation of relatively autonomous AI-DSS without encountering questions of responsibility.

Twitter Hendrik Kempt @hkem

Acknowledgements We thank Fabian Kiessling, Horst Hahn, Helen Heinrichs, Mathias Neugebauer and Bianca Hoffmann for discussions of an early draft. We thank Peter Königs and Jan-Christoph Heilinger for comments, and especially Niël Conradie for valuable feedback on a late draft. Nils Freyer provided valuable support for administrative tasks.

Contributors HK is the main author and guarantor of the paper. SN commented on all parts of the paper.

Funding This research has been funded by Bundesministerium für Bildung und Forschung (BMBF) (German Federal Ministry of Education and Research), Funding Number 01GP1910.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This study does not involve human participants.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Topol EJ. High-Performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25(1):44–56.
2 Molleman E, Broekhuis M, Stoffels R, *et al*. How health care complexity leads to cooperation and affects the autonomy of health care professionals. *Health Care Anal* 2008;16(4):329–41.
3 Mamdani M. Second opinion: is it desirable? *Issues Med Ethics* 1997;5(3):75–7.
4 Payne VL, Singh H, Meyer AN. *Patientinitiated second opinions: systematic review of characteristics and impact on diagnosis, treatment, and satisfaction*. . Elsevier, 2014: 89. 687–96.
5 Shmueli L, Davidovitch N, Pliskin JS, *et al*. Seeking a second medical opinion: composition, reasons and perceived outcomes in Israel. *Isr J Health Policy Res* 2017;6(1):1–11.
6 Van Such M, Lohr R, Beckman T, *et al*. Extent of diagnostic agreement among medical referrals. *J Eval Clin Pract* 2017;23(4):870–4.
7 German Federal Ministry of Health. Info-Blatt 315-02. Krebsfrüherkennungsuntersuchung in derDer gesetzlichen Krankenversicherung (§ 25 Abs. 2 SGB V) / Mammographie. Available: https://www.bundesgesundheitsministerium.de/fileadmin/Dateien/3_Downloads/M/Mammografie/Infoblatt_Krebsfrueherkennung_Mammographie.pdf [Accessed 23 Feb 2021].
8 Puliti D, Duffy SW, Miccinesi G, *et al*. Overdiagnosis in mammographic screening for breast cancer in Europe: a literature review. *J Med Screen* 2012;19 Suppl 1:42–56.
9 Sharma V, Stranieri A, Burstein F, *et al*. Group decision making in health care: a case study of multidisciplinary meetings. *J Decis Syst* 2016;25(sup1):476–85.
10 Carrick P. The Physician's Moral Responsibility. In: *Medical ethics in antiquity*. Springer, 1995: 151–9.

11 Bjerring JC, Busch J. Artificial intelligence and patient-centered decision-making. *Philos Technol* 2021;34(2):349–71.

12 Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare. *J Med Ethics* 2020;46(3):205–11.

13 Frances B. Disagreement. Stanford encyclopedia of philosophy, 2018. Available: https://plato.stanford.edu/entries/disagreement/ [Accessed 11 Nov 2020].

14 Gelfert A, Romanian Academy - Iasi Branch. Who is an Epistemic peer? *Logos Episteme* 2011;2(4):507–14.

15 Kelly T. Peer Disagreement and Higher Order Evidence. In: *Social Epistemology: essential readings*, 2011: 183–270.

16 Weinstein BD. What is an expert? *Theor Med* 1993;14(1):57–73.

17 Luxton DD. Should Watson be consulted for a second opinion? *AMA J Ethics* 2019;21(2):131–7.

18 Castaneda C, Nalley K, Mannion C, *et al*. Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. *J Clin Bioinforma* 2015;5(1):1–16.

19 Braun M, Hummel P, Beck S, *et al*. Primer on an ethics of AI-based decision support systems in the clinic. *J Med Ethics* 2021;47(12):e3.

20 Haenssle HA, Fink C, Schneiderbauer R, *et al*. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann Oncol* 2018;29(8):1836–42.

21 Liu Y, Kohlberger T, Norouzi M, *et al*. Artificial Intelligence-Based breast cancer nodal metastasis detection: insights into the black box for pathologists. *Arch Pathol Lab Med* 2019;143(7):859–68.

22 Steiner DF, MacDonald R, Liu Y, *et al*. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. *Am J Surg Pathol* 2018;42(12):1636–46.

23 Natividad E. Artificial intelligence to give Toronto doctors a 2nd opinion, 2019. Available: https://toronto.citynews.ca/2019/11/04/artificial-intelligence-second-opinion/ [Accessed 11 Nov 2020].

24 Ribeiro MT, Singh S, Guestrin C. "Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016:1135–44.

25 Tjoa E, Guan C. A survey on explainable artificial intelligence (XaI): toward medical XaI. IEEE transactions on neural networks and learning systems, 2020. Available: https://ieeexplore.ieee.org/abstract/document/9233366

26 Lipton ZC. The Mythos of model Interpretability: in machine learning, the concept of Interpretability is both important and slippery. *Queue* 2018;16:31–57.

27 Mittelstadt B, Russell C, Wachter S. Explaining explanations in AI. *Proceedings of the conference on fairness, accountability, and transparency*, 2019:279–88.

28 Doran D, Schulz S, Besold TR. What does explainable AI really mean? a new conceptualization of perspectives. *arXiv* 2017;171000794.

29 Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 2019;1(5):206–15.

30 Wang J, Molina MD, Sundar SS. When expert recommendation contradicts peer opinion: relative social influence of valence, group identity and artificial intelligence. *Comput Human Behav* 2020;107(1):106278.

31 London AJ. Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hastings Cent Rep* 2019;49(1):15–21.

32 Kompa B, Snoek J, Beam AL. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digit Med* 2021;4(1):1–6.

33 Cascella L. Artificial intelligence risks: patient expectations. Med-Pro, 2020. Available: https://www.medpro.com/artificial-intelligence-risks-patientexpectations [Accessed 11 Nov 2020].

34 Weigel S, Kerschke L, Rodriguez-Ruiz A. Künstliche Intelligenz in Ergänzung Zur menschlichen Bewertung mammographischer Auffälligkeiten. *Geburtshilfe Frauenheilkd* 2020;80(06):129.

35 Hawkins A. The world's first robot car death was the result of human error — and it can happen again. The Verge, 2019. Available: https://www.theverge.com/2019/11/20/20973971/uber-self-driving-car-crash-investigation-human-error-results [Accessed 24 Feb 2021].

36 Geis JR, Brady AP, Wu CC, *et al*. Ethics of artificial intelligence in radiology: summary of the joint European and North American multisociety statement. *Can Assoc Radiol J* 2019;70(4):329–34.

37 Patel BN, Rosenberg L, Willcox G. Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ digital medicine* 2019;2(1):1–10.

38 Celi LA, Fine B, Stone DJ. An awakening in medicine: the partnership of humanity and intelligent machines. *Lancet Digit Health* 2019;1(6):e255–7.