AI & SOCIETY (2023) 38:1407–1414
https://doi.org/10.1007/s00146-022-01418-x

OPEN FORUM

"I'm afraid I can't let you do that, Doctor": meaningful disagreements with AI in medical contexts

Hendrik Kempt[1] · Jan-Christoph Heilinger[1] · Saskia K. Nagel[1]

Received: 30 May 2021 / Accepted: 21 February 2022 / Published online: 16 March 2022

Abstract

This paper explores the role and resolution of disagreements between physicians and their diagnostic AI-based decision support systems (DSS). With an ever-growing number of applications for these independently operating diagnostic tools, it becomes less and less clear what a physician ought to do in case their diagnosis is in faultless conflict with the results of the DSS. The consequences of such uncertainty can ultimately lead to effects detrimental to the intended purpose of such machines, e.g. by shifting the burden of proof towards a physician. Thus, we require normative clarity for integrating these machines without affecting established, trusted, and relied upon workflows. In reconstructing different causes of conflicts between physicians and their AI-based tools—inspired by the approach of "meaningful human control" over autonomous systems and the challenges to resolve them—we will delineate normative conditions for "meaningful disagreements". These incorporate the potential of DSS to take on more tasks and outline how the moral responsibility of a physician can be preserved in an increasingly automated clinical work environment.

Keywords Decision support systems · Disagreement · Medical diagnostics · Medical AI · Medical ethics · Pragmatism

1 Introduction

AI-based decision support systems (DSS[1]) are becoming increasingly more sophisticated and integrated, with a growing number of applications for medical and clinical use, especially in diagnostics. With their increased abilities to not only support physicians in making decisions, but to recommend certain actions based on their own decisions independently from physicians, much attention has been given to the requirements for their ethical integration in clinical processes. However, one main issue has so far been given an insufficient amount of attention: the fact that integrated DSS have the potential to cause different kinds of conflicts and disagreements with the expertise of physicians. In reconstructing different types of disagreements between physicians and their AI-based tools and the challenges to resolve them, we will delineate rules about the potential of DSS to take on more tasks and outline how the moral responsibility of a physician can be preserved in a largely automated work environment.

For this, we characterize in Sect. 2 three different types of decision support systems for clinical diagnostics in reference to the measure of meaningful human control (generally introduced by Yu et al (2019), elaborated in Braun et al (2020) for the clinical context). That is, simple DSS that help physicians make diagnostic decisions by taking over non-cognitive processes, advanced DSS which include certain cognitive processes (evidence gathering and first assessments), and fully autonomous DSS that allow for little human intervention points in their diagnostic processes. While we aim to give a theoretical account, we acknowledge that a number of other definitions have been put forward to categorize and operationalize the integration of DSS into clinical contexts. Software as a Medical Device (SaMD) is the term within the legal-regulatory debate (IMDRF 2013), in which question of technical reliability, liability of developers and hospitals, and other questions are evaluated. While

✉ Hendrik Kempt
hendrik.kempt@humtec.rwth-aachen.de

Jan-Christoph Heilinger
jc.heilinger@rwth-aachen.de

Saskia K. Nagel
saskia.nagel@humtec.rwth-aachen.de

[1] Applied Ethics Group, RWTH Aachen University, Theaterplatz 14, 52062 Aachen, Germany

[1] In the following, whenever we use the term "DSS", we refer to AI-based decision support systems.

this debate is far from concluded, the arguments there are often answering different, more specific regulatory questions. Thus, we concentrate our efforts on the ongoing philosophical debate.

The different types of DSS we mention may then lead to different types of conflicts, implying potential shifts in the normative landscape in clinical responsibilities. In Sect. 3, we analyze the challenges that result from integrating conventional or semi-autonomous DSS, and concentrate on the potential conflicts emerging from their integration when taking over more tasks previously performed by human medical personnel.

The difference between conflicts among humans-only on one hand and humans-with-machines on the other will allow us to formulate normative requirements of "meaningful disagreement" (Sect. 4). Having a concept of meaningful disagreement will allow for the further integration of DSS into clinical processes while keeping the normative dimension of a physician's responsibility for the decisions being made at the center of considerations.

Lastly, Sect. 5 will discuss emerging normative questions. The fact that DSS may become ever more autonomous, inevitably creating pressure for physicians to rely on their input in decision-making processes will put forward questions of responsibility and blame in cases of misdiagnoses. Due to the unique role of physicians, attributions of responsibility may remain unchanged, even in a situation of increased interactions with DSS, while the same cannot be said for their blameworthiness.

2 Types of DSS

Different analyses have been put forward categorizing the types of DSS in medical contexts (Topol 2019, Braun et al 2020, for a review, see Sutton et al 2020). This is due to the wide notion of "medical context", as this includes a diverse set of tasks ranging from health file organization over pharmaceutical research to assistance in diagnostics and in risk assessment for surgical decisions. In this paper, we focus on the use of DSS in diagnostics, i.e., the process of matching a patient's symptoms to a specific explanation of physiological or psychological processes. Even in this focused perspective, the use of DSS offers a wide diversity of applications, as the process of diagnosing the symptoms and signs of a patient requires different sub-processes. The sub-processes include correctly identifying the symptoms a patient has that require explanation to gathering further evidence, both ailment- and patient-specific, to settling on a diagnosis (or a set of plausible diagnoses). The decisions necessary in these different sub-processes can most be aided by different kinds of DSS with different levels of sophistication and autonomy.

To systematize different types of DSS in the diagnostic process, we follow Yu et al. (2019) and Braun et al (2020) in distinguishing three types of DSS according to the level of "meaningful human control". Accordingly, they distinguish between conventional AI-driven support systems (C-AI), integrative support systems (I-AI) and fully autonomous ones (F-AI). This approach remains generally agnostic about the specific capacities of a DSS, but assesses whether human interactors or supervisors have relevant access to the decision-making process to assert control. This human control is intended to develop DSS alongside the ethical requirements of clinical decision-making, rather than adjusting the latter to account for the developments of the former.

Specifying this distinction for diagnostic contexts, examples for C-AI may be a chatbot guiding the anamnesis of patients, or an algorithm helping to take and sort the relevant medical images of a patient, which we call knowledge-based evidence (Sutton et al 2020). Such DSS are already in development or in use, like the anamnesis-chatbot "Ana" (Denecke et al 2018), and will be further improved in the future, but it seems unlikely that their 'autonomy' will increase much further, if not connected to other DSS. For I-AI, we may consider diagnostic recommender systems that largely operate on their own but do not make decisions, or algorithms that gather non-knowledge-based evidence, such as in radiomics. As Braun and colleagues point out, integrative systems like this may still change certain clinical norms and thereby require careful guidance in their integration and reflections on the potential changes of norms in clinical contexts.

Lastly, a fully autonomous evidence-gathering, diagnosis-producing, and treatment-recommending AI may count as what Braun and colleagues call the F-AI. As we intend to keep our analysis tethered to disagreements potentially occurring in the not-too-distant future, we leave out considerations of fully autonomous AI in clinical contexts and instead focus on I-AI, from which current trends and developments can be extrapolated. F-AI machines are too uncertain in their realization to engage with normative consequences that reach beyond mere speculation. Additionally, those assumed technologies usually do not allow for productive conclusions about rules of disagreement for physicians (see Wilkinson et al (2020) for further elaboration on the often misguided promises of "precision medicine"), which will be the point of interest for this paper.

3 Conflicts with DSS: a terminological proposal

The ever more integrated use of conventional and integrative DSS in many different contexts has generated a prolific scholarly description and analysis of the interactions occurring between humans and these machines. Some describe these interactions coming from an instrumental perspective as essentially advanced tool-use (Köhler 2020), in which the behavior of the machine is merely a highly complex process that can be adjusted to respond to complex human actions. Others have taken the autonomous processes of these machines to amount to something more action-like, rendering the description of these interactions more akin to cooperation, collaborations or partnerships (e.g., Patel et al. (2019) for diagnostic imaging, see Nyholm (2018) for a general analysis).

Without a coherent terminology to describe these interactions and the specifics of the relation between human and DSS, it will be difficult to provide an accurate description of the frictions that may arise within these interactions. Additionally, the structure of these frictions will depend on features of the DSS, requiring a technology-sensitive approach to assess the conflicts occurring between humans and a specific system. The wide variety of potential applications as well as the particularly high stakes of conflicts in using DSS in medical contexts presuppose clear conceptual distinctions for an accurate analysis. We propose some terminological choices that allow disentanglement of some conceptual and normative confusions, and to propose some norms to help integrate DSS further into clinical processes.

For this, we first suggest understanding any friction between the expected and the actual output between human physicians and DSS as *conflicts*. This means, that the diagnostic output of a DSS, and the diagnostic prediction (or expectation) of a physician, are incongruent with each other.

Importantly, we propose to understand conflicts not so much as a problem to be overcome but as an indication of an opportunity to be seized. This positive reframing is based on the idea that certain conflicts, disagreements and ambiguities are indicative of differences that need to be acknowledged and offer an opportunity for improvements. On this account, (the right kind of) conflict is the place where actual progress can be made, progress in epistemological, practical or other terms.

The idea to positively reframe conflicts is indebted to the tradition of American pragmatism, notably the work of John Dewey, whom we quote here at length (Dewey 1922, 301):

> What is to be done with […] facts of disharmony and conflict? After we have discovered the place and consequences of conflict in nature, we have still to discover its place and working in human need and thought. What is its office, its function, its *possibility*, or use? In general, the answer is simple. Conflict is the gadfly of thought. It stirs us to observation and memory. It instigates to invention. It shocks us out of sheep-like passivity, and sets us at noting and contriving. Not that it always effects this result; but that conflict is a *sine qua non* of reflection and ingenuity. When this possibility of making use of conflict has once been noted, it is possible to utilize it systematically to substitute the arbitration of mind for that of brutal attack and brute collapse.

Thus, Dewey proposes to see conflicts in human need and thought as an extension of conflict in nature. Both instigate innovative change and progress.[2] Here, we offer a further extension of this idea to cover conflicts between humans and machines and will show in the following how an analysis of conflicts elucidates opportunities for improvements in clinical processes.

3.1 Mistakes and malfunctions

Conflicts between physicians and DSS can be caused by different forms of friction or incongruence. Diagnostics is known to operate with wide margins of uncertainty, as both our knowledge about diseases in general as well as their formation in individual patients is still limited. Thus, proposing different diagnoses can be based on three different reasons: either, the physicians is wrong in their assessment, the DSS is wrong in theirs, or both are within range of probable results.

To clarify how we should go about dealing with these different causes for conflict, we first analyze the two instances in which either side commits an error. Thus, a conflict caused by such an error can and should be resolved by taking the error-free side.

These two types concern the incorrect operation by one of the interacting parties, the physician or the DSS. If a physician makes a mistake in operating the system or in their own deliberations, the DSS's recommendations may be in conflict with the anticipated results. The same counts for a physician who misunderstands or misapplies diagnostic evidence or theories and thereby ends up with erroneous beliefs about the diagnosis, while the DSS may be of faultless assistance. On the other hand, an incorrectly operating system may be operated correctly but still produce false results. In the former case we speak of *mistakes* (where the error lies with the physician), in the latter of *malfunctions* (where the error lies with the system). One could also speak of mistakes as "operator errors", and malfunctions as "operating errors".

However, due to the complexity and autonomy of decision-making procedures within these systems, the descriptions of human–machine interactions as more than mere mistakes or malfunctions ought to be accounted for in the reconstruction of conflicts. With DSS that, as some philosophers have argued, ought to be capable of replicating the human decision-making process (Lin 2015), especially in cooperating with humans (Nyholm 2018), the frictions potentially occurring here can gain a new quality.

[2] Cf. also Heilinger (2016; 2020), ch. 3.

3.2 Disagreements

This leads us to the third option, in which a conflict is caused by either both sides being wrong, or both sides having sufficient error-free evidence to support their incompatible claims. Physicians may, for example, begin perceiving some DSS as quasi-agents due to their complex problem-solving skills and the need for physicians to rely on their correct operation. Insofar as it may be impossible for the physicians to fully reduce the process that led a DSS to generate a recommendation to its deductive elements, they may begin to perceive a conflict between a DSS and their own opinion as one about equally weighty beliefs. A physician cannot simply reject a DSS suggestion anymore, even if they came to different results. Conflicts about beliefs or actions with this kind of similar standing, however, are usually considered a different kind of conflict, i.e., *disagreements*.

Yet, the *perception* or *presumption* of a disagreement is, per se, not sufficient for establishing the presence of actual disagreement. In human–human disagreements, one side (or both) may be mistaken in their argument, and thereby not be in disagreement with each other but in a mistake-based conflict (malfunctions-based conflicts may not occur in human–human disagreements).

As a diagnostic AI system only detects patterns, it does not "know" what these patterns represent nor is it designed to provide interpretation. In consequence, disagreeing with the conclusion of a machine may be entirely based on the machine's arbitrary misinterpretation of certain features of an image. To call such a conflict between a physician and an AI a proper disagreement seems implausible and undesirable: Implausible, because it suggests we can disagree with (quasi-) agents who use unintelligible processes, if any at all. We equally do not disagree with a parrot that merely repeats what it was trained to say without representing the meaning of the sentence. And it is undesirable, because it shifts the burden of proof partially towards the physician who has to justify their decision in case of conflict, which potentially solely rests on a machine's lack of representations and "common sense." For these types of conflicts, we would suggest counting them among malfunctions (Flach 2019).

However, this reconstruction does not catch the full scope of which I-AI is capable, as not all of these conflicts can be satisfyingly understood as caused by malfunctions or mistakes. Reconstructing a semi-autonomous diagnostic recommendation system as a simple tool denies the far-reaching role and influence it can have in the diagnostic process, in which the AI may actually perform most of the cognitive work. The system's degree of autonomy, even if not at the level of full human autonomy, and its ability to assess certain features on its own to guide the procedure clearly surpasses what usually is expected from and contributed by "tools". The relevance of the cognitive work, the ability of some DSS to access non-knowledge-based evidence, and the altogether insufficient explainability (Mittelstadt et al. 2019) gives credence to the perceptions of those conflicts as disagreements.

Perceived disagreements thus are often irresolvable in the way disagreements among physicians are usually resolved: Discursive methods of exchanging reasons to change an opinion under current C-AI and I-AI methods are unachievable due to the fundamental difference in "reasoning" of deep neural networks and other machine-learning paradigms (Pelaccia et al. 2019).

An example-case may help to clarify the proposed terminological set-up: Imagine a physician utilizing a diagnostic machine for assisting in breast-cancer detection as a second opinion, one of the more common uses of DSS in clinical contexts. For their first patient, the physician writes down wrong information from the patient's file about previous cancer-treatments in their own assessment. After the patient went through the DSS-assisted analysis, the DSS correctly refers to previous treatment methods, while the physician's diagnosis planned first-time treatments. This conflict is caused by a mistake, as the physician failed in their duty to carefully evaluate the patient. For their second patient, the physician proposes their tentative diagnosis and awaits the machine's response. As it turns out, the DSS claims the patient to be in the very late stages of cancer with very little chance of healing. As the patient does not report any kind of issue, it is soon discovered that the machine misread the medical images provided due to a malfunction. For the third patient, both physician and machine have different diagnoses, with the physician diagnosing the patient to not have breast cancer, while the DSS is positively diagnosing an early stage of the disease. While the physician has drawn a different conclusion, they can appreciate how the DSS reached a different result (and are sufficiently uncertain to outright reject such result). While in all three situations, both physician and DSS disagree, it requires further analysis as to how they are in conflict and why their respective resolution causes different ethical problems.

From the perspective of a physician in a clinic relying on largely independently operating AI, ultimately, perceiving the interaction with the machine as a cooperation will, in case of conflict, often lead to *perceived disagreements*. This constitutes a morally distinctive and sensible situation, for it is not inconceivable, then, that physicians will feel the burden of proof shifting towards them, with their expertise questioned, when disagreeing with the diagnostic recommendation of a DSS. Instead of improving the physician's diagnosis, such perceived conflict puts their own expertise under pressure of justification. A certain "shadow expertise" is established through the DSS, in which physicians may be in potential competition with a machine that cannot take responsibility for its operational error, while, in fact, physicians must take this responsibility.

3.3 From meaningless to meaningful disagreements

With an idea in place about the potential for some sort of disagreement among human–machine interactions in medical diagnostics, the analysis can turn towards the normative problem of how to integrate these DSS in medical, and in particular diagnostic, processes without compromising their quality. As these issues are concerning the integration into physicians' work environments, we focus on two problematic consequences of integrating non-human expertise in clinical processes. This section will describe them and show how they motivate the need for something we propose to call "meaningful disagreement". Such a term can alleviate these consequences and stands in contrast to "meaningless disagreements". These are, according to our terminology, merely perceived disagreements which fall back into the categories of mistakes or malfunctions.

A *first* problematic consequence of integrating DSS in diagnostic processes can be an increased reluctance of doctors to make their own, independent diagnoses (Grote and Berens 2020), as the correctness of these diagnoses can be challenged by AI. Some have even suggested (Kompa et al. 2021; Mozannar and Sontag 2020) that recommendations of an AI may be tailor-made to fit the success score of individual physicians. This proposal claims that the benchmarks when an AI proposes a diagnosis to a physician should be set in comparison to performances of individual physicians. In these cases, an AI only proposes a diagnosis if its confidence score is on average significantly higher for diagnosing a certain disease over the success score of the physician to whom it is recommending the diagnosis. This shifts the burden of proof substantially towards the physician.

As such a score will negatively affect the willingness to propose more "risky" diagnoses, i.e., those of comparatively rare diseases, it will potentially decrease the overall diagnostic precision of physicians who use DSS (Grote and Berens 2020, 208). This turns the very idea of assisting physicians to make more precise diagnoses on its head. The risk and the associated moral costs of being wrong will increase for the physician if there is an I-AI proposing an alternative that the physician cannot classically "reason" with (ibid.).

A *second* problematic consequence of integrating DSS in diagnostic processes comes from the complementary perspective of conflict-avoidance: rather than reluctance to propose one's own genuine diagnosis, we can expect the misuse of these DSS to generate diagnoses a physician can then agree to without ever having developed and proposed their own diagnosis. Notably, DSS are conceived as *support* systems, i.e., not certified to make their own decisions. However, they are capable of decreasing human control to a binary "confirming" or "denying" the suggestion made by a DSS. This conceivably exerts a pressure to "let the machine go first" (McDougall (2019), stresses this point from a patient-perspective) and adjust one's own diagnosis accordingly, so that no disagreement occurs. In doing so, the machine's prowess de facto replaces human expertise.

Unfortunately, Braun et al., despite discussing the need for meaningful human control, reject the idea that such perceived disagreements may be a potential area of conflict (Braun et al 2020, 7). We, however, argue that we should account for those. We can encounter I-AI DSS that are so complex that those interacting with them have at least pragmatic reasons to take a DSS's diagnosis as something they can disagree with. Braun and colleagues consider conflicting diagnoses a "misnomer" because for most decisional situations, there is not one precise way to move forward (ibid.), and because most diagnoses are judgments of relative (un-) certainty, relative to the given evidence and available resources and theories. However, as analyses have suggested in other areas of human–machine interactions (HMI), the perceptions of humans in these interactions factor in the expectations of norms regulating these interactions (Bankins and Formosa 2019). We take these perceived disagreements to be not only a psychological fact but a challenge to be incorporated into the norms of a clinic.

4 Confronting meaningful disagreements

So far, we have advanced a terminological proposal and listed normative requirements for conflicts of experts with DSS in diagnostic settings and found that those requirements are insufficiently met by considering them as mere operating or operational errors. Depending on the type of DSS used, their integration in clinical processes and the work they are performing, the conflicts emerging in the diagnostic process between a machine and a physician can vary considerably in quality. Thus, a normative analysis of these different types of conflicts requires clarity about how humans, in our case diagnostic physicians, can *meaningfully disagree* with DSS—without the pressure of being blamed if they disagree while being wrong. Only if this is secured can physicians continue to fully exercise their expertise. With such analysis in place, its results may also be generalized and apply to other contexts in which reliance on and interactions with DSS play an important role.

Following the preconditions of human–human disagreement (as studied in philosophy, cf. e.g. Christensen 2007 and see 3.2), meaningful disagreement requires the acknowledgment of the reasonableness on the side of both disagreeing parties and one's own epistemic limitations. A meaningful disagreement among humans may not occur if one of the two sides does not provide reasons (by merely rejecting the request or being unable to deliver). Questions of taste, for example, may not be resolved by exchanging reasons, as these reasons are purely subjective and personal and thus inaccessible for reasoned debate. Thus, if one side claims something to be true while refusing to give any explanation for their position, we may not disagree with them in any meaningful way, as the disagreement cannot be resolved satisfactorily or without someone changing their minds without being given reason to do so. A meaningful disagreement, while not necessarily solvable, is at least debatable.

If we transfer this condition to human–DSS conflicts, the disagreement can be meaningful only if the disagreeing person can understand how the machine's suggested diagnosis came about. A machine’s proposition (even a negative one, like the rejection of a diagnosis based on a lack of conclusive evidence derived from procedural standards), may count as conceivable and thereby as a meaningful disagreement if the physician can read an evidential assessment into the diagnosis. Returning to our example about using DSS as assistance in breast-cancer detection: a perceived disagreement was one in which the physician had no reason to doubt a DSS’s proposed diagnoses except for that fact that the physician proposed a different one. A meaningful disagreement, in this case, is one in which the interpretability of the DSS’s results is sophisticated enough that the physician can appreciate the “reasoning” of the machine.

However, the issue of the burden of proof may lie within those diagnoses that are suggested by a DSS based on evidence unavailable to the physicians, like the aforementioned non knowledge-based systems. The burden-of-proof-concern, as formulated with Braun et al. (2020) as well, problematizes the possibility that a physician’s expertise will be measured equally against a machine’s rather considering the latter an addition to the former. The difference in explanations, or rather the lack of explanations on the machine’s diagnosis, will make this shift possible if merely success scores will be compared.

This worry seems rather theoretical for two reasons: first, the use of such evidence, e.g., radiomics, presupposes a particular expectation of a physician in the first place. As long as the physician instructs the I-AI to analyze medical images with non-knowledge-based methods (instead of an F-AI doing this all on their own), the physician should expect a variety of results (even ones the physician may not think are likely). Pragmatically, then, even surprising results from a DSS can, in the case of conflicting assessments, count as a meaningful disagreement if the use of a DSS to assess evidence was intentionally done.

Second, for some diagnoses, non-knowledge-based evidence is essential. Radiomics is used precisely because physicians have no other way of producing evidence at this stage and are thereby not capable of proposing an opinion the machine could disagree with. Usually, the results from radiomics are merely conducive to forming an opinion in the first place.

Furthermore, if someone in a human–human disagreement about a decision unilaterally decides to break the tie and proceed, we usually are comfortable to hold the deciding side responsible for the success or failure of the decision. However, in human–DSS disagreements, no such equivalency is present: the human physician ought to remain the sole decision-maker (Geis et al. 2019), since responsibility can be assured only here. In these cases, it is the “prerogative to be wrong” of a physician should they decide falsely against the DSS. DSS, being uninterested and unaffected by wrong decisions, cannot be held responsible and thereby ought not to have the last word in a decision. Especially in the medical context, the assumed responsibility of physicians for guarding the diagnostic and therapeutic process should increase the expectations for any other entity to take responsibility in this process (including engineers who program DSS, for an overview of responsibility in medical contexts at large see Rogers 2020).

Any diagnostic suggestion made by DSS ought to have a certain baseline-reasonableness and physicians ought to be able to break a tie between their diagnosis and the machine’s. From this fact, it follows that a physician ought to also have reasons when they disagree with a DSS’s diagnosis. Analogous to human–human disagreements, a physician may not merely discard someone else’s suggested diagnosis if such diagnosis exhibits a baseline of reasonableness. This mirrors the perception of DSS as quasi-agents without attributing agency to them, as the physician is merely epistemically obligated to take the DSS-based diagnosis into consideration. Yet, rejecting a proposed DSS’s diagnosis without being able to provide adequate reasons to do so, is as problematic as rejecting a second opinion from a fellow physician.

While this argument does shift the burden of justification towards the physician, it does so on ethical grounds, not on epistemic ones: rejecting a DSS-proposed diagnosis should require justification even if the physician is willing to take full responsibility if the rejected diagnosis turns out to be correct. The mere fact that physicians ought to remain in charge does not justify their rejection of evidence, which a well-designed DSS certainly is able to provide. This shifts the obligation to an essentially ethical demand to consider DSS-diagnoses even in case of disagreements. The key point in justifying decisions in disagreements, then, is not a physician’s epistemic authority, but their moral responsibility (for the discussion of medical expertise see e.g., Cassam (2017) and Applbaum (2017)).

5 Reducing responsibility?

However, one important remaining issue is the consideration of blame and responsibility in human–machine interaction of medical diagnostics. While a physician clearly ought to be in charge to remain responsible for misdiagnoses based on a disagreement with DSS, there seems to be an intuitive reduction of blame (if not responsibility) if such a disagreement does not occur: If a DSS and a physician are in agreement about the diagnosis, and no operator or operational errors have influenced decision-making process, yet the diagnosis turns out to be wrong, the amount of blame the physician deserves may decrease.

To strengthen this point from a different perspective, consider the following: If two physicians are wrong in their diagnosis, while one has consulted a DSS for guidance and the other has not, it seems plausible to assign less blame to the one who did include a DSS in the decision-making process. The former physician clearly followed their duties to incorporate as many perspectives as a clinical process allows, while the latter refused to do so.

The discussion surrounding reduced, shared, or collective responsibility in human–machine interaction is well underway and too wide to cover here (see Nyholm 2018; Köhler 2020 for discussions). However, these debates are often centered around fundamental relational problems of human agency and responsibility. In clinical contexts, these fundamental questions are often replaced by pragmatic, legally required role-responsibilities. Physicians take a special role in the healing process, rendering discussions around shared responsibility in clinical contexts often fruitless. However, if we can still hold a physician responsible for a misdiagnosis even when they used the best available medical knowledge and technological resources, we can ask whether this role-responsibility is still an appropriate source for the moral attitudes of blame, i.e., they may not be blameworthy but still responsible.

Hence, instead of blame, we may praise physicians incorporating DSS into their clinical processes as long as the integration is according to legally binding and ethically justified guidelines. Only this way, we can assume that physicians aim in using the best available medical knowledge and technology to correctly diagnose a patient, even at the risk of having to make a judgment call in a meaningful disagreement with a DSS.

6 Conclusion

The diagnostic process inevitably includes some degree of uncertainty, a fact that will remain unchanged for the foreseeable future. Integrating autonomous technologies such as DSS can significantly improve the overall diagnostic precision by decreasing the physician’s workload, by providing evidence that helps come to conclusions, and by presenting diagnostic suggestions on their own. However, the remaining uncertainty will indubitably lead to conflicts between physicians and these technologies and their capacities.

To avoid the risks of shadow expertise and irresolvable disagreements, we aimed to provide both a terminological proposal and a normative analysis of disagreements. This allowed reconstructing certain requirements to the decision-making process of physicians that accounted for the role DSS can play in improving diagnosis, while not shifting the burden of justification towards physicians. Physicians remain the addresses of responsibility, and can, therefore, reject recommendations of DSS if they disagree. However, it does not seem plausible that physicians alone should still remain fully blameworthy for a misdiagnosis when that diagnosis has been a result of physician–DSS interactions.

And while this investigation of integrating DSS in clinical processes has focused on the relationship between physicians and those systems, several perspectives remain in need of further elaboration. Ever more independent systems will affect physician–patient relationships, require more careful reassessments of liability approaches, and put into question how we center the health of patients in an increasingly digitized clinical workflow.

Acknowledgements We would like to thank the organizers and participants of the “Embedding AI in Society” conference for the opportunity to present and discuss an early draft. Further thanks belong to Niël Conradie, Peter Königs, Fabian Kießling, Horst Hahn, Bianca Hoffmann, Markus Wenzel, and Helen Heinrichs for discussions on some key concepts. We thank Nils Freyer for administrative assistance. Lastly, we like to thank the reviewers and editors of the journal for their helpful comments and leading the process.

Funding Open Access funding enabled and organized by Projekt DEAL. This work has been developed and funded by the project ELSA-AID (Grant Number 01GP1910A) of the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung (BMBF)), as well as the project KIPeriOP (Grant Number 2520DAT10G) by the German Federal Ministry of Health (Bundesgesundheitsministerium (BMG)).

Declarations

Conflict of interest We have no conflicts of interests to declare.

References

Applbaum AI (2017) The idea of legitimate authority in the practice of medicine. AMA J Ethics 19(2):207–213

Bankins S, Formosa P (2019) When AI meets PC: exploring the implications of workplace social robots and a human-robot psychological contract. Eur J Work Organ Psychol 29(2):215–229

Braun M, Hummel P, Beck S, Dabrock P (2020) Primer on an ethics of AI-based decision support systems in the clinic. J Med Eth. https://doi.org/10.1136/medethics-2019-105860

Cassam Q (2017) Diagnostic error, overconfidence and self-knowledge. Palgrave Commun 3:17025. https://doi.org/10.1057/palcomms.2017.25

Christensen D (2007) Epistemology of disagreement: the good news. Philos Rev 116:187–218

Denecke K, Hochreutner SL, Pöpel A, May R (2018) Talking to ana: a mobile self-anamnesis application with conversational user interface. In: Proceedings of the 2018 international conference on digital health, pp 85–89

Dewey J (1922) Human nature and conduct. An introduction to social psychology. Holt, New York

Flach P (2019) Performance evaluation in machine learning: the good, the bad, the ugly, and the way forward. Proc AAAI Conf Artif Intell 33(1):9808–9814

Geis JR, Brady AP, Wu CC, Spencer J, Ranschaert E, Jaremko JL, Kohli M et al (2019) Ethics of artificial intelligence in radiology: summary of the joint European and North American multisociety statement. Can Assoc Radiol J 70(4):329–334

Grote T, Berens P (2020) On the ethics of algorithmic decision-making in healthcare. J Med Eth 46(3):205–211

Heilinger J-C (2016) Konflikte in Der Ethik. Anmerkungen aus pragmatistischer Perspektive. In: Nida-Rümelin J, Heilinger J-C (eds) Moral, Wissenschaft Und Wahrheit. de Gruyter, Berlin, pp 145–159

Heilinger J-C (2020) Cosmopolitan responsibility. Global injustice, relational equality, and individual agency. de Gruyter, Berlin

International Medical Device Regulators Forum-IMDRF (2013) Software as a Medical Device (SaMD): Key Definitions. https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-131209-samd-key-definitions-140901.pdf

Köhler S (2020) Instrumental robots. Sci Eng Ethics 26(6):3121–3141

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

Lin P (2015) Why ethics matters for autonomous cars. Autonomes Fahren. Springer, Berlin, pp 69–85

McDougall RJ (2019) Computer knows best? The need for value-flexibility in medical AI. J Med Eth 45(3):156–160

Mittelstadt B, Russel C, Wachter S (2019) Explaining Explanations in AI. In: Proceedings of the conference on fairness, accountability, and transparency, pp 279–288

Mozannar H, Sontag D (2020) Consistent estimators for learning to defer to an expert. In: International conference on machine learning. PMLR, pp 7076–7087

Nyholm S (2018) Attributing agency to automated systems: reflections on human-robot collaborations and responsibility-loci. Sci Eng Ethics 24:1201–1219. https://doi.org/10.1007/s11948-017-9943-x

Patel BN, Rosenberg L, Willcox G et al (2019) Human–machine partnership with artificial intelligence for chest radiograph diagnosis. NPJ Digit Med 2:111. https://doi.org/10.1038/s41746-019-0189-7

Pelaccia T, Forestier G, Wemmert C (2019) Deconstructing the diagnostic reasoning of human versus artificial intelligence. CMAJ 191(48):1332–1335

Rogers W (2020) Moral responsibility in medicine: where are the boundaries? Lancet. https://doi.org/10.1016/S0140-6736(20)31643-3

Sutton RT, Pincock D, Baumgart DC et al (2020) An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 3:17. https://doi.org/10.1038/s41746-020-0221-y

Topol EJ (2019) High-performance medicine: the convergence of human and artificial intelligence. Nat Med 25(1):44–56

Wilkinson J, Arnold KF, Murray EJ et al (2020) Time to reality check the promises of machine learning-powered precision medicine. Lancet Digit Health 2(12):677–680. https://doi.org/10.1016/S2589-7500(20)30200-4

Yu K, Beam AL, Kohane IS (2019) artificial intelligence in healthcare. Nat Biomed Eng 2(10):719–731