Agree to disagree: The symmetry of burden of proof in human-AI collaboration[1]

KR Jongsma[1] & M Sand[2]

[1] Medical Humanities, University Medical Center Utrecht, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, PO Box 85500, 3508 GA Utrecht. Email: k.r.jongsma@umcutrecht.nl

[2] Department of Values, Technology and Innovation, Faculty of Technology, Policy and Management, Jaffalaan 5, 2628 BX Delft, The Netherlands. Email: m.sand@tudelft.nl

In their paper "*Responsibility, second opinions and peer-disagreement: ethical and epistemological challenges of using AI in clinical diagnostic contexts*", Kempt and Nagel discuss the use of medical AI systems and the resulting need for second opinions by human physicians, when physicians and AI disagree, which they call the rule of disagreement (RoD).[1] The authors defend RoD based on three premises: First, they argue that in cases of disagreement in medical practice, there is an increased burden of proof (better to be conceived as a *burden for justification*) for the physician in charge, to defend why the opposing view is adopted or overridden. This burden for justification can be understood as an increased responsibility. In contrast, such burden does allegedly not arise, when physicians agree in their judgment. Second, in those medical contexts where humans collaborate with humans such justification can be provided, since human experts can discuss the evidence and reasons that have led them to their judgment, through which the sources of disagreement can be found and a justified decision can be made by the physician in charge. Third, unlike human-to-human collaboration, such communicative exchange is not possible with an AI system. Due to AI's opacity, the physician in charge has no means of illuminating why the AI disagrees. Conclusively, the authors propose RoD as a solution. RoD suggests that a second human expert should be consulted to advise in cases of human-AI disagreement. Once AI systems become more widespread in clinical practice, it can be expected that such type of disagreement occurs more frequently. AI, after all, is being implemented, because it promises, amongst others, higher accuracy, which implies that some abnormalities will be detected that the physician would have missed.[2] Hence, it is laudable to discuss the moral implications of disagreement for clinical practice and consider whether these cases are analogous to those of in which human experts disagree. In the following, we will focus in particular on the first premise of the argument, consider whether the stated asymmetry between agreement and disagreement indeed holds and what the implications are for RoD. We propose a more refined idea of medical expertise and we outline some concerns regarding the efficiency of medical AI, if RoD were adopted.

The authors' view on disagreement is best expressed in the following paragraph: *"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."*
This view is strongly impacted by the idea that physicians are experts and agreement between the initial judgement and second opinion ought to increase their confidence levels in the rightness of the

---

[1] This manuscript is a pre-print version of a commentary that was published in the *Journal of Medical Ethics*. Please cite the original publication: **Jongsma KR, Sand M. Agree to disagree: the symmetry of burden of proof in human–AI collaboration. Journal of Medical Ethics. 2022;48(4):230-1. doi: 10.1136/medethics-2022-108242**. Both authors contributed equally to the paper.

diagnosis. The reverse happens in cases of disagreement: Confidence in the rightness of the diagnosis should decline, if they disagree and, additional justificatory weight (the weight of the burden of proof, as they call it), ends up on the shoulder of the physician in charge. First, this picture neglects the fact that the very step of requesting a second opinion already requires a justification as to why confidence has been low, why a specific colleague has been chosen for consultation and what gains and insights have emerged from this collaboration. These are justificatory burdens that are on surface not at all less weighty in cases of agreement than they are in the cases of disagreement. Second, while having more confidence in a medical judgment might in general be justified in cases of agreement, each individual case of agreement requires a separate justification: a reason for reaching the same conclusion. Without understanding these reasons, we cannot know whether coming to the same conclusion should indeed be called agreement, or must rather be seen as an effect of the employment of different heuristics by both physicians, as conclusions emerging from different reasons or even just mere coincidence. Through the process of exchanging reasons medical professionals examine each other's levels of expertise and quality of judgment and, thereby, validate each other's expertise. In those cases that induce the need for collaboration, expertise has to be continuously reestablished through reason exchange both in cases of agreement and disagreement. Expertise without such exchange of reason is a label that might suffice in public settings, where non-experts might have good reasons to frequently rely on expert knowledge, but cannot suffice when a particular medical situation leads to uncertainty and poses epistemic problems with far-reaching consequences. In short, expertise in medical practice should not be understood as a label that guarantees epistemic certainty. So, contra Kempt and Nagel, even in cases of agreement, medical professionals must be wary that their judgement and that of their colleagues is fallible.

We largely concur with the author's second premise that the burden for justification poses a challenge for human-AI collaboration, because medical AI systems are typically opaque, meaning that they are inscrutable for humans.[3] If it is true – as established before – that in collaborative settings, justificatory burdens arise both in cases of agreement and disagreement, we see now that neither can be sufficiently fulfilled in human-AI collaboration; physicians cannot understand or explain the AI and, therefore, cannot identify why an AI system came to agree or disagree with the initial judgement. The severity of this explanatory problem varies depending on the context. Arguably, in relatively simple diagnostic AI applications, where link uncertainty – the relation between the actual phenomenon in question and the features of this phenomenon that the AI model uses to predict its development e.g. – is low (e.g. skin cancer diagnosis based on visual assessment), algorithmic opacity seems less concerning.[4] Yet, the responsibility in cases of agreement and disagreement is the same: If physicians do not want to naively have their views confirmed, they have to justify why they consider the AI systems' output as a confirmation.

The symmetry of responsibility viz. burden for justification for cases of agreement and disagreement that we defend has implications for the RoD of human-AI systems as proposed by Kempt and Nagel. RoD suggests that "*[i]f 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."* [1] We believe that it follows from the symmetry of agreement and disagreement that the requirement of second opinion must also be applied to cases of agreement. This means that in cases of agreement and disagreement of humans and AI-systems it is required that another physician considers the case at hand and provide a second (or rather third) opinion. If – as RoD suggests – in cases of disagreement an additional second opinion ought to be considered, and if it were true – as we argued before – that even many, if not all cases, of agreement also require living up to a substantive burden for justification, we must assume that none of the expected efficiency gains of AI employment will materialize.[5] To the contrary, the introduction of AI systems, would make diagnostic processes even more time-consuming. This might pose a considerable reason to reject RoD.

Aside from RoD, there are various other ways to tackle the problems of human-AI agreement and disagreement. There is a vast amount of literature that suggests ways to making AI more understandable and interpretable [6,7], so that physicians can in fact compare and assess their own reasons in light of the evidence considered by the AI-system. Further, one might focus on forward-looking responsibility to prevent the explanatory and consequently responsibility gaps and develop institutionalized solutions located not only in the context of clinical practice (such as RoD), but in the broader ecosystem that this innovation is.[8] One of these suggestions is to educate physicians and doctors to raise awareness of physicians of their grown oversight responsibilities, to keep in check whether the AI systems declines in accuracy, whether they are still safe in terms of personal data protection, and whether the data that is being fed is of reasonable quality to name but a few.[2] Forward-looking responsibilities would advance doctors capabilities to collaborate with medical AI, both in cases of disagreement and agreement.

References

1. Kempt H, Nagel SK. Responsibility, second opinions and peer-disagreement: ethical and epistemological challenges of using AI in clinical diagnostic contextsJournal of Medical Ethics Published Online First: 14 December 2021. doi: 10.1136/medethics-2021-107440
2. Sand M, Durán JM, Jongsma KR. Responsibility beyond design: Physicians' requirements for ethical medical AI. Bioethics. 2022;36(2):162-9. doi: https://doi.org/10.1111/bioe.12887.
3. Durán JM, Jongsma KR. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. J Med Ethics 2021;47:329-335.
4. Sullivan E. Understanding from Machine Learning Models. The British Journal for the Philosophy of Science. 2020: doi: 10.1093/bjps/axz035.
5. Topol EJ. Deep Medicine - How Artificial Intelligence Can Make Healthcare Human Again. New York: Basic Books; 2019
6. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, the Precise Qc. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Medical Informatics and Decision Making. 2020;20(1):310.
7. 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.
8. Santoni de Sio F, Mecacci G. Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. Philosophy & Technology. 2021;34(4):1057-84. doi: 10.1007/s13347-021-00450-x.