

Responsibility beyond the black box: a medical student's reading of Giubilini's paper "It is not about AI, it's about humans"

Anna-Katharina Große

Angaben zur Veröffentlichung / Publication details:

Große, Anna-Katharina. 2026. "Responsibility beyond the black box: a medical student's reading of Giubilini's paper 'It is not about AI, it's about humans'." *Journal of Bioethical Inquiry* 23 (3): 775–77. <https://doi.org/10.1007/s11673-026-10566-4>.

Nutzungsbedingungen / Terms of use:

CC BY 4.0





Responsibility Beyond the Black Box: A Medical Student's Reading of Giubilini's Paper "It Is Not About AI, It's About Humans"

Anna-Katharina Große 

Received: 3 March 2026 / Accepted: 18 March 2026 / Published online: 20 May 2026
© The Author(s) 2026

The title of Alberto Giubilini's article, "It is not about AI, it's about humans. Responsibility Gaps and Medical AI" (2025) suggests that the ethical challenges of artificial intelligence (AI) in medicine are less about the technical nature of algorithms and more about our evolving understanding of human responsibility. In this commentary, I reflect on Giubilini's arguments by situating them within a concrete clinical encounter from my medical training in cardiology.

In a rather casual conversation, my supervising cardiologist remarked that I might be among the last generations of medical students to learn auscultation properly, as this skill would increasingly be taken over by AI systems. This shift is reflected in recent studies on AI-based stethoscopes showing that algorithms can identify certain valvular heart diseases (VHD) with high diagnostic accuracy, in some settings even rivalling the performance of experienced clinicians (Zhou et al. 2025).

At first, this appeared to be merely a description of an ongoing technological evolution. But the more I reflected on it, the clearer it became that it pointed to a deeper ethical question: What happens to responsibility when we rely on "black-box" systems that may prioritize predictive performances over explainability (Kiener 2022; London 2019)? Or, put more sharply:

Is understanding a necessary condition for responsibility? This question points to the core of the "responsibility gap" debate, where the loss of predictive control over learning automata is often seen as a barrier to ascribing moral agency (Matthias 2004).

I first approached Giubilini's text guided by the sceptical intuition that opacity and a loss of control necessarily undermine responsibility. Against the backdrop of these concerns, Giubilini's article offers a necessary conceptual shift. His aim is not to assess the technical performance of AI but to provide a framework for how responsibility remains meaningfully distributed within healthcare contexts involving AI. A central claim is that AI systems cannot be moral agents: they lack the capacity for moral reasons and thus cannot be blameworthy. This clarifies that AI does not remove responsibility from humans but forces us to redistribute it. This redistribution addresses the challenge of epistemic opacity, often described as the "black-box" nature of AI in clinical decision-making (London 2019). Rather than arguing that such opacity does not undermine responsibility, Giubilini treats this as one possible assumption and explores its implications alongside the contrasting view that opacity can, in fact, erode responsibility. To illustrate this first assumption, we can draw on London's (2019) argument that medical decision-making has always involved elements that are not fully transparent, including human judgment shaped by experience and heuristics, and yet physicians remain responsible for their decisions.

A.-K. Große (✉)
Medical Faculty, University of Augsburg, Stenglingstraße
2, 86156 Augsburg, Germany
e-mail: anna-katharina.grosse@uni-a.de

This perspective redirected my thinking: the ethical core is not the opacity itself but how it is managed professionally. Applied to my cardiology example, this suggests that even if I cannot explain how an AI system classifies heart sounds, I may still bear responsibility for relying on its output. This justification rests on an active choice to trust a tool based on its verified clinical validity rather than a complete understanding of its code.

At this point, Giubilini's distinction between different dimensions of responsibility becomes particularly helpful. Drawing on Shoemaker (2011), he distinguishes between attributability, answerability, and accountability.

First, regarding attributability: if the AI system were to misclassify a diastolic murmur, such as an aortic insufficiency, this technogenic error cannot be attributed to the doctor as a moral character flaw. Because the process within the black box is inaccessible, its proximal cause lies in the system's internal functioning rather than the physician's negligence (Matthias 2004; Santoni de Sio and Mecacci 2021).

However, the physician remains answerable to the patient and must be able to justify the decision to use the AI. This justification does not rest on understanding the algorithm itself but on the tool's verified reliability in clinical practice. In this sense, the physician bears responsibility for what Fischer and Ravizza (1998) would call the "normative bridge" between the technical output and the clinical decision.

Extending Giubilini's framework beyond individual clinicians, accountability shifts to the institutional level. If medical training, for example, at universities deprioritizes manual auscultation in favour of AI, as my supervisor predicted, the accountability for resulting risks lies with the system. This can be understood by applying Danaher's concept of "tragic choice" to institutional contexts, where institutions are confronted with situations that require balancing competing moral goods (Danaher 2022). In the case of auscultation, this tragic choice concerns the question of whether to prioritize superior diagnostic accuracy through AI-supported analysis or the preservation of traditional bedside skills.

This move reveals that the "responsibility gap" is located at the systemic level, where curricula define the professional identity of the next generation. This raises a fundamental normative question: What counts as medical competence under such conditions

and which forms of responsibility are future physicians expected to develop? In this sense, AI functions as a mirror: it reflects the ways in which responsibility is structured, distributed, and sometimes obscured within healthcare systems.

In retrospect, I read the article less as an attempt to resolve ethical concerns about AI and more as an invitation to reflect on the nature of human responsibility itself. My initial intuition that AI might undermine responsibility has not disappeared, but it has shifted. It is now directed less at individual clinicians and more at the systems within which responsibility is shaped, taught, and distributed.

Before closing, I would like to add one final thought. Whether or not we are currently facing a genuine responsibility gap, I believe we have a duty to prevent such gaps from emerging or widening in the future. This requires more than technical regulation; it calls for ongoing reflection on our relationship with AI and on the healthcare systems in which we operate. Even if AI systems cannot enter into relationships with us, we enter into relationships with them through the ways we rely on them and allow them to shape our healthcare systems and professional identities (Santoni de Sio and Mecacci 2021). Remaining attentive to how these relationships reshape our understanding of responsibility at a professional and systematic level may be one of the most important ethical practices for future physicians and healthcare systems (Kerasidou et al. 2022).

Acknowledgements This commentary was produced as part of my assignment for the Oxford Bioethics Winter School 2026, hosted by the Ethox Centre, University of Oxford. I would like to thank the organisers Mehrunisha Suleiman (University of Oxford), Suzanne Metselaar (VUmc Amsterdam), Gerben Meynen (VU Amsterdam), and Ruth Horn (University of Augsburg). Special thanks go to Alberto Giubilini, who gave me the opportunity to comment on his paper in a "critic-meets-author" dialogue and provided me with feedback on my commentary, and to my supervisor Ruth Horn for supporting me in this endeavour.

Author's Contribution Anna-Katharina Große is the sole author of this commentary and is responsible for its content.

Funding Open Access funding enabled and organized by Projekt DEAL. This manuscript was produced as part of the Oxford Bioethics Winter School funded by the Ethox Centre Oxford. The author received a scholarship to participate in the Winter School by the Institute for Ethics and History of Health in Society and the Department for Medical Education at the University of Augsburg. Open access publication fees

are covered by the University of Augsburg as part of Project DEAL.

Data Availability Not applicable.

Declarations

Ethics Approval Not applicable. This commentary does not involve human participants or animals, and ethical approval was not required.

Consent Not applicable.

Competing interests The author declares no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Danaher, J. 2022. Tragic choices and the virtue of technoresponsibility gaps. *Philosophy & Technology* 35(2): 26.
- Fischer, J. M., and M. Ravizza. 1998. *Responsibility and control: A theory of moral responsibility*. Cambridge University Press.
- Giubilini, A. 2025. It is not about AI, it's about humans: Responsibility gaps and medical AI. *Journal of Bioethical Inquiry* 22(3): 527–537.
- Kerasidou, C.X., A. Kerasidou, M. Buscher, and S. Wilkinson. 2022. Before and beyond trust: reliance in medical AI. *Journal of Medical Ethics* 48(11): 852–856.
- Kiener, M. 2022. Can we bridge AI's responsibility gap at will? *Ethical Theory and Moral Practice* 25: 575–593.
- London, A.J. 2019. Artificial intelligence and black-box medical decisions: Accuracy versus explainability. *Hastings Center Report* 49(1): 15–21.
- Matthias, A. 2004. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology* 6(3): 175–183.
- Santoni de Sio, F., and G. Mecacci. 2021. Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology* 34(4): 1057–1084.
- Shoemaker, D. 2011. Attributability, answerability, and accountability: Toward a wider theory of moral responsibility. *Ethics* 121(3): 602–632.
- Zhou, Z., K. Xie, Y. Huang, et al. 2025. Automatic diagnosis of left valvular heart disease based on artificial intelligence stethoscope. *JACC: Advances* 4(11): 101993.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.