Human Computation (2026) 12:1:1-10

ISSN: 2330-8001, DOI: 10.15346/hc.v12i1.149

The Misguided Expectations of Human Overseers in AI in Healthcare

ROANNE VAN VOORST, UNIVERSITY OF AMSTERDAM

ABSTRACT

This commentary proposes an idea based on the outcomes of collaborative workshops and ethnographic inquiry within hospital settings, exploring the dynamic interplay between medical practitioners (clinicians and nurses) and artificial intelligence (AI). The research reveals a poignant finding: the prevailing emphasis on ethical AI places undue strain on physicians, obligating them to engage in continuous 'digital literacy' training. This imposition not only exacerbates the existing burdens of healthcare professionals but also fosters a misguided sense of security, given their non-specialist status in software programming and AI comprehension. The investigation underscores the intricate challenges and ethical quandaries inherent in the human-AI partnership within the domain of healthcare. Furthermore, the notion of physicians as the 'human overseer,' regarded as a requisite component of 'ethical AI' per legislative mandates, is revealed to be somewhat fallacious, shifting a complex ethical dilemma towards individual responsibility, as not all clinicians in this loop possess the capacity to rebut AI outcomes or grasp the complexities of AI algorithms.

1. INTRODUCTION

Across the globe, governments are championing big data and AI systems as essential to the future of healthcare: they are considered potential solutions to the anticipated healthcare crisis. At the same time, research consistently highlights that the datasets driving these AI systems often (re)produce social biases, leading to discrimination and infringing upon personal autonomy. Automated decision-making systems, including those increasingly utilized in healthcare, tend to adversely affect the poor and middle-class by implementing mechanisms of control (Eubanks, 2018; Passchier, 2021). This raises critical questions about the balance between automated intelligence and human judgment in healthcare. Understandably, then, and in response to these concerning findings, there has been a surge of public and academic scrutiny regarding algorithmic ethics. In recognition of the risks associated with AI, both governments and technology companies have developed numerous legal frameworks and regulatory guidelines for the usage of AI in healthcare, aimed at promoting ‘ethical,’ ‘responsible,’ or ‘fair’ AI—now numbering in the hundreds globally.

While these regulations differ in specific content, they all emphasize the importance of explainability in algorithms, ensuring that software programmers and, ideally, end-users can grasp how decisions are made. In the realm of healthcare, these end-users are professional caretakers: the hundreds of thousands clinicians and nurses who are increasingly working with AI. Their oversight as human operators is deemed a vital component of labeling AI systems as 'ethical'.

Maintaining human involvement in the decision-making process is essential for protecting human rights (Wagner, 2019). Enarsson et al (2022) note that human oversighters seem to have become a standard solution for solving the issues of transparency, bias, legal security and systemic risks relating to automation (149). They explain that "keeping a human in the loop is a deliberate attempt to maintain human agency and accountability, and to provide legal safeguards and quality control. Hybrid decision-making can thus be said to operate in-between somewhat counterbalancing ambitions, where the wish for effectivization and automation may require a reduction of human discretion at the same time as legal requirements of maintaining human oversight and agency may necessitate such discretion" (2001, p. 124).

This viewpoint is further supported by the Council of Europe Expert Group on Internet Intermediaries (MSI-NET), which emphasizes that assigning significant decision-making roles to humans is crucial for the protection of human right (see also Wagner, 2019, p. 108). Moreover, in the EU's Artificial Intelligence Act, one paragraph (14, in article 113) is specifically dedicated to the need for human oversight in high-risk AI utilization: "Human oversight shall aim to prevent or minimise the risks to health, safety or fundamental rights that may emerge when a high-risk AI system is used in accordance with its intended purpose or under conditions of reasonably foreseeable misuse." In order to do so, in the same paragraph it is explained that natural persons overseeing high-risk AI systems must understand their capabilities and limitations to monitor performance and identify anomalies. They need awareness of potential biases in AI output, particularly when providing information or recommendations. As stated in articele 14 of the Artificial Intelligence Act, individuals must also be able to interpret AI results, override outputs, and intervene or halt the system as needed.

In line with these warnings, this commentary argues that such expectations of caretakers to take up this job effectively is often unrealistic, as well as it is unfair to a profession already struggling to keep with the high demands of working in public healthcare. I do, by no means, aim to eliminate the role of human oversight in the context of healthcare; indeed, doing so would imply that AI could make autonomous decisions—a scenario that is obviously undesirable. However, I do want to point out that the current requirement for human oversight in ethical AI in healthcare is equally problematic, and that the issue of AI oversight by natural persons in the realm has received insufficient attention in public and academic discussions about what 'ethical AI' truly entails in the daily lives and work of the clinicians and nurses who interact with it. This criticism aligns, however, with conclusions of other scholars working on AI, as I elaborate below.

Recent literature indicates that, while the regulatory requirements for human oversight assume that humans must and can help mitigate some risks associated with AI systems, their ability to effectively do this will depend on various factors, including the types of systems they are overseeing, the transparency of those systems, and their roles and working conditions, such as the training they receive (Enqvist, 2023).

Ben Wagner (2019) has studied three fields in which, he argues, human agency in decision-making is currently debatable, as humans only have nominal control or responsibility for decisions: self-driving cars, border searches based on passenger name records, and content moderation on social media. He concludes that there exist a vast number of cases in which significant automation is actually taking place, as long as somewhere in the process a human is still perceived to maintain oversight, which offers a façade of humane control: he refers to such nominal control as 'rubber-stamping' automation: "Existing legal rules that, for example, forbid or allow certain forms of automation do so on the assumption that a 'human in the loop' means that an actual human 'check' will take place of the results of the automated system. If the person is able to only rubber-stamp the results produced by the algorithm, then these systems should perhaps more accurately be called 'quasi-automated.' This is particularly the case when the company involved spends little time or energy ensuring that staff are properly trained or prepared to make these decisions, or that they have sufficient time to make the decision themselves" (2019, p. 114).

Hence, in order to understand when humans are really able to oversee AI, we need to go beyond what is written in policies and draw attention to how they unfold in the everyday life, in the workspaces where humans increasingly collaborate with AI; Enarsson et al (2022) point to the need for research into hybrid decision-making environments to go beyond legal doctrinal studies, by the implementation of a socio-technical perspective and the use of empirical studies.

My research, grounded in empirical and extensive work in hospitals, offers such a case example. In this commentary I contend that there are two main reasons why we cannot, and should not expect too much of professional-caretakers-as-overseers-of-AI: first, literature debates have established that the notion that any human can act as an autonomous overseer of AI is outdated; instead, decision-making in human-AI interactions should be perceived as a hybrid functioning system. Second, coming from the research that I conducted with my team in six hospitals around the world over the past years: physicians often lack both the time and the necessary training to adequately fulfil the requirements for effective human oversight. We are only halfway in the research (it runs until 2026), yet our studies already consistently indicate that many doctors do not feel comfortable being honest with management or public about the challenges and concern that exist around their utilization of AI. Consequently, as I will argue in this text, this creates a false sense of security within and beyond the institutions, and shifts a complex ethical dilemma towards individual responsibility, and away from hybrid functioning systems.

After providing a brief background of the project, I discuss both issues and relate our findings to relevant work of other scholars. I end with a call for more empirical research into the daily dynamics of hybrid human-AI systems.

2. CONTEXT AND RESEARCH SETTING

Anthropological fieldwork formed the base of the research project of which I am the Principal Investigator and in which I collaborate with a team of two PhDs, a Postdoctoral researcher and several research assistants. Together, we are in the process of conducting a five-year, international ethnographic study supported by the European Research Committee, which investigates the challenges of digitization in healthcare and the ethical complexities of human-AI collaboration. The project spans diverse hospital settings across the Netherlands, China, Norway, Estonia, Denmark and the United Arab Emirates, utilizing focus group interviews and ethnographic fieldwork to illuminate the lived experiences of healthcare practitioners.

Findings presented in this commentary are mainly based on an extensive series of interviews and roundtable discussions involving 121 healthcare professionals—including doctors and nurses—alongside 35 ethicists and software engineers. These discussions were integral to the research project.

Some examples of the types of AI systems part of this study may be helpful to help ground the arguments developed in this piece: in two hospitals, we follow clinicians working with an AI tool that draws the organs of a body affected by cancer, into a 3D visual. In this digital drawing it estimates the cancer-affected area, information that is useful for clinicians deciding on treatment plans. In yet another hospital an AI tool that assists nurses in adjusting insulin doses for diabetes patients, based on real-time glucose levels and nutritional data. Although the AI systems we made part of our study differ greatly, they have in common that they involve everyday human-nonhuman collaboration and decision-making. The research specifically focuses on clinicians and nurses who collaborate with AI; our focus is not on the AI system and its technologies, but much more on the humans that work with them, more specifically on the ways in which the humans in our study make decisions together with AI. The research does not compare cases but rather contrasts them as a means to sharpen our understanding of how humans and nonhumans (AI) collaborate, and with which potential results for public healthcare.

3. OVERWORKED, AND THEN MORE

A second problem underlying the false sense of security that is currently constructed, is the lack of effective AI training for caretakers. To clarify: the issue is not that doctors and nurses are not offered AI-related training or that they are unwilling to participate. The real problem is that both

the training programs and the healthcare professionals who attend them cannot keep pace with the rapid advancements in AI.

Many of our interlocutors complained about the ongoing offering of new, not-to-be-missed AI trainings in their hospital wards. They explained they were already overworked, having to deal with full waitingrooms and neverending lists of patients, even without any extra digital training added to their to-do lists. As a consequence, in practice, they often engage with AI-trainings in a half-hearted manner or complete courses while feeling that they do not adequately understand the material. For instance, one physician remarked during a workshop that he participates in the training purely to check off requirements imposed by management due to the AI system purchased for his ward. Another doctor compared his experience of struggling through yet another AI training to attending a Zoom meeting: "You participate a little, answer some emails, and occasionally check your social media." This might sound blasé, but even the many professional caretakers who genuinely tried to follow each training offered to them in full concentration indicated that they felt unsure about their knowledge, afterwards—they are specialized in medicine, after all, not programming, so by far not all of them are able to grasp the workings of the AI systems they are supposed to oversee.

The sentiments expressed by these workshop attendees appear to be representative of a broader trend. Other scholars have noted that the assumption that AI should be "explainable" or transparant to doctors is naive. For example, in a concerning article in The Lancet, Ghassemi, Oakden-Raymer, and Beam (2021) argue that current explainability methods cannot provide clear and reliable explanations for each individual decision made by the AI system. Hence, the expectations placed upon professional caretakers to keep up with AI developments, seems unrealistic and even unfair, considering their often already heavy workload. By uncritically assuming that healthcare professionals can act as independent overseers of AI, we create a false sense of security that does not exist in reality. We also shift the responsibility for identifying the crucial human actors in the decision-making process—ranging from programmers to physicians—disproportionately onto the caregivers, many of whom are ill-equipped to bear that responsibility. Of course, there are exceptions: some interlocutors felt that they truly understood what they were dealing with in their collaboration with AI, and this understanding was, to the best of our outsiders' assessments, accurate. In fact, in the hospitals where we conduct fieldwork, we are following some doctors who are themselves coding, or who closely collaborate with developers and technology companies to co-design new algorithms. I have written elsewhere about the latter group (Van Voorst, 2024). However, the problem of the false sense of security that can prevail in an entire hospital ward still persists in such occasions: colleagues who have a less comprehensive understanding of the technology often feel pressured by management or more tech-savvy staff to use it anyway. Such negative effects, or stigmatising interactions, where AI technologies are not embraced, raises important micro-level interactionist questions around the pressures that some people may experience towards using technologies despite misgivings (Brown and Meyer, 2015).

Furthermore, even among those who believe they have a good grasp of how AI works, we still know too little about how AI technology influences human decision-making processes. Indeed, the evidence indicating that individuals actively intervene or resist AI technologies is shockingly limited (Hannah-Moffat, 2013; Monahan and Skeem, 2016; Peeters, 2020). Many people may be unwittingly swayed by its perceived impartiality, or—as already touched upon—might simply be under significant time constraints (or other pressures) that prevent them from consistently verifying AI-generated outcomes. It is relevant to point out here that the inclination to rely on AI solutions tends to increase in situations where managers and professionals feel more vulnerable. Consequently, their ability to engage in reflexive thinking and resistance is compromised, while their need to trust in these systems intensifies (Brown, 2021). This is especially true in environments where "digital artifacts and infrastructures have been framed as urgent and essential" (Pickersgill, 2020, p. 16).

4. CONCLUDING REMARKS AND CALL FOR FURTHER RESEARCH

In this commentary, I have proposed that while the demand for human oversight in AI systems is understandable and even crucial—especially in the context of healthcare where the idea of fully autonomous AI decision-making is widely regarded as problematic—there are also significant risks associated with the concept of the human overseer that must be discussed.

A recurring concern has emerged, both in my own research as well in that of aforementioned colleagues: the often unrealistic expectations placed upon medical professionals to possess a comprehensive understanding of algorithmic technologies. As these technologies proliferate in healthcare, practitioners are expected to function as effective overseers of algorithmic decision-making, a role deemed essential for the ethical deployment of AI. This expectation rests on the flawed assumption that all nurses and clinicians can seamlessly, or even with effort, interpret the calculations or recommendations generated by AI systems and make informed decisions about whether to adhere to or diverge from such advice. Not everyone in the medical realm has the talent or ability to interpret statistics, or understand how code is built.

Digitizing trainings, although already offered, are currently not always able to solve this problem, specifically not as many healthcare professionals are not provided with sufficient extra time to follow such trainings in their packed schedules. The rising emphasis on ethical AI imposes an additional burden on healthcare professionals, compelling them to undertake continuous training in 'digital literacy' or what we could call AI literacy. While well-intentioned, this expectation exacerbates the already significant pressures faced by medical practitioners, many of whom lack backgrounds in software programming or AI. Consequently, the reliance on physicians and nurses to oversee AI functionality generates a false sense of security regarding the ethical deployment of these technologies. Hence, I contend with Wagner that offering human oversighters more time, both for understanding AI and for reflecting on its outcomes during the decision-making process,

is crucial (see Wagner, 2018, p. 115, point 1). I would suggest that AI developers and hospital management should advocate for and facilitate tailored training programs that fit within the schedules of healthcare professionals. These training programs should be more interactive than informative. They should not only focus on AI literacy or explain how the human-nonhuman decision-making technically works for this tool, but also provide practical examples relevant to their daily tasks, ensuring that professionals feel empowered to engage with AI technology. Importantly, examples of errors or potential biases must be standardly included in the trainings—we found that this is not always, nor everywhere the case.

Furthermore, the prevailing notion that clinicians serve as the essential human overseer, in a chain of 'humans in the loop' deemed indispensable for ensuring ethical AI as dictated by legislative frameworks, warrants scrutiny. This framework incorrectly shifts complex ethical responsibilities onto individuals who may not possess the expertise to critically assess AI outputs or fully grasp the intricacies of the algorithms they are working with. Let us not forget that, next to the liability for decision-making about patients, clinicians and nurses could now also risk carrying the responsibility for ambiguities and biases that may already been written into the code. Moreover, they are sometimes expected to become an expert in how to interpret AI results—an expertise which cannot be quickly developed by everyone, and certainly not by everyone working in a healthcare system under immense pressure and while being expected to also be as time efficient as possible in their daily practice. Ideally, nurses and clinicians work in a culture where healthcare professionals feel safe expressing concerns and asking questions about AI systems. In practice, however, we found that often they do not, as they are afraid to be judged as stupid, or old fashioned (see also reference withheld 2025). Regular meetings with external parties about AI adoption, and anonymous surveys might also facilitate sharing of experiences and feedback. The encouragement by management of informal mentorship relationships where experienced AI professionals assist caretakers in understanding and using AI in the specific work context could also be an option, just as peer discussion groups for sharing practical tips and best practices. But again: for this, healthcare professionals need to be provided with time: time to adapt, to process, and to reflect.

It is essential to explore these pressing issues in future research and debates, as reevaluating the responsibilities assigned to medical professionals in an increasingly AI-driven healthcare landscape is crucial. Further research, involving empirical cases where humans collaborate with AI, can help lay bare what happens on the ground, thus pushing our thinking away from what was expected from human overseers, based on regulatory frameworks and policies. In cases where researchers find that such expectations are unrealistic, amendments can be made—both to the policies, and to the workpractice. This does require a gradual, iterative implementation of AI in workspaces; a need that seems to go against the current speed of AI adoption.

It is always easier to find criticism, than to come up with solutions, and I do not believe the suggestions above solve all problems. But by continuing to address these concerns, we can foster a more realistic approach to integrating AI into clinical practice, ultimately safeguarding both

patient well-being and ethical standards in healthcare.To address these challenges, future research should focus on closely examining the outcomes of regulations and expectations in real-world settings. It is imperative to track the interactions between humans and machines, particularly in the daily work context of healthcare professionals. Empirical work, including observations and in-depth interviews, is most suitable for this aim. Additionally, we must maintain a realistic perspective in public and scholarly debates on what can be expected from physicians, nurses, and other practitioners who make daily decisions in collaboration with AI. Their decisions must be regarded as the results of hybrid decision-making, rather than as the result of human oversight alone. And, importantly, stakeholders involved in these hybrid processes need to be able to dare and speak out about their collaboration with AI systems, and the extent to which AI impacts what they think and decide. This is a first step towards fostering an environment in which these professionals will also feel safe to voice their concerns when regulations or expectations are not grounded in reality, or whether they lose grip about who decides what, and why. Such research and debates will be vital for creating an effective and ethically sound partnership between humans and AI in healthcare.

5. REFERENCES

Amoore, L. (2020). *Cloud ethics: Algorithms and the attributes of ourselves and others*. Duke University Press.
Artificial Intelligence Act. (n.d.). *Article 14*. Retrieved June 13, 2025, from https://artificialintelligenceact.eu/article/14/

Brown, P. (2021). *On vulnerability: a critical introduction*. Routledge.

Brown, P. R., & Meyer, S. B. (2015). Dependency, trust and choice? Examining agency and 'forced options' within secondary-healthcare contexts. *Current Sociology*, *63*(5), 729-745. https://doi.org/10.1177/0011392115590

Callon, M., & Law, J. (1995). Agency and the hybrid collectif. *South Atlantic Quarterly*, 94(2), 481-507. https://doi.org/10.1215/00382876-94-2-481

De Togni, G., Erikainen, S., Chan, S., & Cunningham-Burley, S. (2021). What makes AI 'intelligent'and 'caring'? Exploring affect and relationality across three sites of intelligence and care. *Social Science & Medicine*, 277, 113874. https://doi.org/10.1016/j.socscimed.2021.113874

Enarsson, T., Enqvist, L., & Naarttijärvi, M. (2021). Approaching the human in the loop – legal perspectives on hybrid human/algorithmic decision-making in three contexts. *Information & Communications Technology Law*, *31*(1), 123–153. https://doi.org/10.1080/13600834.2021.1958860

Enqvist, L. (2023). 'Human oversight'in the EU artificial intelligence act: what, when and by whom?. *Law, Innovation and Technology*, *15*(2), 508-535. https://doi.org/10.1080/17579961.2023.2245683

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11), e745-e750.

Hannah-Moffat, K. (2013). Actuarial sentencing: An "unsettled" proposition. *Justice quarterly*, 30(2), 270-296. https://doi.org/10.1080/07418825.2012.682603

Hayles, N. K. (2022). Ethics for cognitive assemblages: Who's in charge here?. In S. Herbrechter (Ed.), *Palgrave handbook of critical posthumanism* (pp. 1195-1223). Cham: Springer International Publishing.

Monahan, J., & Skeem, J. L. (2016). Risk assessment in criminal sentencing. *Annual review of clinical psychology*, 12(1), 489-513. https://doi.org/10.1146/annurev-clinpsy-021815-092945

Passchier, R. (2021). *Artificiële intelligentie en de rechtsstaat: over verschuivende overheidsmacht, Big Tech en de noodzaak van constitutioneel onderhoud*. Boom Publishers.

Peeters, R. (2020). The agency of algorithms: Understanding human-algorithm interaction in administrative decision-making. *Information Polity*, 25(4), 507-522. https://doi.org/10.3233/IP-200253

Pickersgill, M. (2020). Uncertainty work as ontological negotiation: adjudicating access to therapy in clinical psychology. *Sociology of Health & Illness*, 42, 84-98. https://doi.org/10.1111/1467-9566.13029

Savolainen, L., & Ruckenstein, M. (2024). Dimensions of autonomy in human–algorithm relations. *New Media & Society*, 26(6), 3472-3490. https://doi.org/10.1177/14614448221100802

van Voorst, R. (2024). The medical tech facilitator: an emerging position in Dutch public healthcare and their tinkering practices. *Medicine Anthropology Theory*, *11*(2), 1-23. https://doi.org/10.17157/mat.11.2.7794

van Voorst, R. (2025). Redefining intelligence: collaborative tinkering of healthcare professionals and algorithms as hybrid entity in public healthcare decision-making. *AI & SOCIETY*, 40, 3237-3248. https://doi.org/10.1007/s00146-024-02177-7

Wagner, B. (2019), Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems. *Policy & Internet*, 11(1), 104-122. https://doi.org/10.1002/poi3.198

Zanzotto, F. M. (2019). Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research*, 64, 243-252. https://doi.org/10.1613/jair.1.11345