A Feminist View of Medical AI Harm

Clair Baleshta, Western University (Canada)
International Conference on Computer Ethics: Philosophical Enquiry (CEPE) 2023, Chicago, IL

Keywords: artificial intelligence, harm, medicine, feminist philosophy, autonomy

1. Introduction

Despite widespread concerns about the harms of Artificial Intelligence (AI) in contexts like medicine, there has been little explicit analysis of what constitutes a 'harm' when looking at the impact of these AI systems (Altman et al. 2018, Smuha 2021). Such an analysis is necessary, however, for both the accurate identification and the effective reduction of harms imposed by this technology. This is a larger task than it may appear to be. The difficulty of providing a clear guiding concept of harm can be seen through an appeal to the philosophical literature – despite harm playing a central role in moral, legal, and political philosophy, there is still significant disagreement regarding how to define the concept (Shiffrin 2012). Nevertheless, this paper aims to develop a distinctive, feminist account of AI harm, connecting such harm to a view rooted in relational theories of well-being. Given the primacy of well-being concerns in medicine (Crisp 2021), the account will be particularly valuable for understanding the harms of AI in the medical context, as well as clarifying the existing literature surrounding this harm. However, it will also be applicable to, and have implications for, our understanding of AI harm more generally.

2. Medical AI and Harm

AI systems, which encompass a broad range of computational approaches to solving complex problems, are widely regarded as some of the most influential technology shaping our future (Schubbach 2021). In medicine specifically, AI is expected to offer substantial improvements in areas such as clinical diagnosis and treatment, patient safety, and administrative optimization (McCradden et al. 2020, Aggarwal et al. 2022). While the use of this technology will clearly offer significant benefits, its potential to cause serious harm has also been widely recognized (Altman et al. 2018, Aggarwal et al. 2022). Many of these concerns surround the potential for AI systems to infringe on privacy, contain biases that lead to inaccurate outputs, and leave patients without explanations for diagnoses and treatment decisions (Sparrow and Hatherley 2019).

However, despite the increasing attention to AI harm in areas like medicine, current approaches to regulating AI lack and explicit analysis as to what constitutes 'harm' when looking at the impact of this technology (Altman et al. 2018, Smuha 2021). Without this foundation, efforts to identify and address the adverse effects of these systems are bound to result in inconsistencies, inaccuracies, or gaps. For instance, current legal assessments of AI risk analysis focus primarily on AI's adverse impact on individuals, overlooking the fact that these systems can also cause collective and societal harms that are distinct from individual harms (Smuha 2021). By neglecting this societal dimension, regulations meant to tackle harms caused by AI are often not suitable to address these wider harms and protect social interests (Smuha 2021). Similarly, marginalized groups are at a heightened risk when it comes to medical AI, and may experience different types or degrees of harm than other patients (Sparrow and Hatherley 2019). Without

a clear concept of harm that can account for these variations, current approaches to addressing medical AI harm may overlook the harms to those most vulnerable.

3. Defining Harm

As such, it is clear that an explicit conceptual definition of harm is needed in the medical AI ethics literature. However, supplying such a definition is no easy task. Drawing on the extensive philosophical literature on harm, one can see that that the concept is still highly contested (Shiffrin 2012, Bradley 2012). Among other things, debates arise regarding whether harm should be understood in a comparative or non-comparative sense (Feinberg 1984, Shiffrin 2012), whether it should be viewed as a state or an event (Thomson 2011, Hanser 2008), and even whether we should retain a philosophical concept of harm to begin with (Bradley 2012).

When it comes to understanding AI harm, however, a larger problem with existing accounts may lie outside of these debates. In particular, these accounts often ignore the ways in which social and systemic factors are relevant to our understanding of harm (Dea 2020, Miller 2022), resulting in individualistic conceptions that overlook differences in vulnerability. In other words, standard philosophical accounts of harm are themselves subject to many of the same issues that this paper hopes to address, and as such are ill-fit to account for AI harm. Given that harm is an important tool for understanding the negative impacts of medical AI, we must therefore work toward developing a suitable definition of AI harm.

4. A Feminist View of Medical AI Harm

This paper aims to propose such a definition, introducing an account that defines harm in terms of adverse impacts on well-being, where well-being is understood in specific sense. To begin motivating this account, I engage with the existing harm literature to demonstrate that well-being is already an underlying element in many accounts of harm (Feinberg 1984, Shiffrin 2012, Harman 2009, Hanser 2008). This justification also involves appealing to recent work from Johansson and Risberg (forthcoming), which argues that standard approaches to harm do not take serious enough the centrality of well-being (3).

I then argue that a more precise understanding of well-being is required when it comes to medical AI harm. Though well-being is itself a central concept in philosophy, it is often understood in highly idealized ways (Knowles 2018, 69). As such, this traditional view is inadequate when it comes to addressing the impact that social factors often have on well-being (Ginzberg 1991, Knowles 2018). Based on this, I instead advocate for adopting a feminist understanding of the concept. One major feature of feminist approaches to well-being is their relational view of the self – a view which focuses on the impact that social and structural relationships have on agents (Miller 2022).

This account will support the identification of medical AI harms by directing attention to the ways this technology may negatively impact patients' well-being. As well-being is understood in a relational sense, it will be particularly attuned to the disparate impact of medical AI on members of marginalized groups. Moreover, the view is unified in that it can also be applied to the collective and societal harms of these systems. In this way, my account will allow for an understanding of medical AI harm that is not structured around individual harms exclusively. Finally, given that the account identifies how the harms of medical

AI are intertwined with social factors, it will have implications for our understanding of responsibility with respect to these harms.

References

Aggarwal, N., M. E. Matheny, C. Shachar, S.Y. Wang, and S. Thadaney-Israni. 2022. "Artificial Intelligence in Healthcare." In The Oxford Handbook of AI Governance, edited by Justin Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew Young, and Baobao Zhang. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780197579329.013.50.

Altman, M., A. Wood, and E. Vayena. 2018. "A Harm-Reduction Framework for Algorithmic Fairness." IEEE Security Privacy 16 (3): 34–45. https://doi.org/10.1109/MSP.2018.2701149.

Bradley, B. 2012. "Doing Away with Harm." Philosophy and Phenomenological Research 85 (2): 390–412.

Crisp, R. 2021. "Well-Being." In The Stanford Encyclopedia of Philosophy, edited by Edward N. Zalta, Winter 2021. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2021/entries/well-being/.

Dea, S. 2020. 2020. "Toward a Philosophy of Harm Reduction." Health Care Analysis 28 (4): 302–13. https://doi.org/10.1007/s10728-020-00405-x. Feinberg, J. 1987. Harm to Others. Oxford University Press.

Ginzberg, R. 1991. "Philosophy Is Not a Luxury." In Feminist Ethics, edited by Claudia Card. Lawrence: University Press of Kansas.

Hanser, M. 2008. "The Metaphysics of Harm." Philosophy and Phenomenological Research 77 (2): 421–50.

Harman, E. 2009. "Harming as Causing Harm." In Harming Future Persons: Ethics, Genetics and the Nonidentity Problem, edited by Melinda A. Roberts and David T. Wasserman, 137–54. International Library of Ethics, Law, and the New Medicine. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-1-4020-5697-0_7.

Johansson, J., and O. Risberg. Forthcoming. "A Simple Analysis of Harm." Ergo: An Open Access Journal of Philosophy. Accessed January 9, 2023. https://philarchive.org/rec/JOHASA-13.

Knowles, C. 2018. "Feminist Perspectives on Well-Being." In The Routledge Handbook of Well- Being, edited by Kathleen Galvin. London, UK: Routledge.

McCradden, M. D., A. Baba, A. Saha, S. Ahmad, K. Boparai, P. Fadaiefard, and M. D.

Cusimano. 2020. “Ethical Concerns around Use of Artificial Intelligence in Health Care Research from the Perspective of Patients with Meningioma, Caregivers and Health Care Providers: A Qualitative Study.” Canadian Medical Association Open Access Journal 8 (1): E90–95. https://doi.org/10.9778/cmajo.20190151.

Miller, S. C. 2022. “Toward a Relational Theory of Harm: On the Ethical Implications of Childhood Psychological Abuse.” Journal of Global Ethics 18 (1): 15–31. https://doi.org/10.1080/17449626.2022.2053562.

Schubbach, A. 2021. “Judging Machines: Philosophical Aspects of Deep Learning.” Synthese 198 (2): 1807–27. https://doi.org/10.1007/s11229-019-02167-z.

Shiffrin, S. V. 2012. “Harm and Its Moral Significance.” Legal Theory 18 (3): 357–98. https://doi.org/10.1017/S1352325212000080.

Smuha, N. A. 2021. “Beyond the Individual: Governing AI’s Societal Harm.” Internet Policy Review 10 (3). https://doi.org/10.14763/2021.3.1574.

Sparrow, R., and J. Hatherley. 2019. “The Promise and Perils of AI in Medicine.” International Journal of Chinese & Comparative Philosophy of Medicine 17 (2): 79–109. https://doi.org/10.24112/ijccpm.171678.

Thomson, J. J. 2011. “More on The Metaphysics of Harm.” Philosophy and Phenomenological Research 82 (2): 436–58.