




# MedAIcine: A Pilot Project on the Social and Ethical Aspects of AI in Medical Imaging

Sophie Jörg<sup>1</sup> (✉) , Paula Ziethmann<sup>2</sup> , and Svenja Breuer<sup>3</sup> 

<sup>1</sup> Munich School of Philosophy, Kaulbachstraße 31/33, 80539 Munich, Germany  
sophie.joerg@hfph.de

<sup>2</sup> University of Augsburg, Universitätsstraße 2, 86159 Augsburg, Germany

<sup>3</sup> Technical University of Munich, Augustenstraße 46, 80333 Munich, Germany

**Abstract.** As artificial intelligence continues to advance and permeate various aspects of our lives, it is crucial that we consider the ethical and social implications of these developments. With its pilot project ‘MedAIcine’ the newly founded Center for Responsible AI Technologies (‘CReAITe’) strives for critically reflecting vital concepts and conflicts regarding the responsible design and use of AI in medical imaging, using an interdisciplinary approach called ‘embedded ethics and social science’. Drawing on perspectives of developers, physicians, and patients across three different use cases (radiology, endoscopy, and dermatology), we identify key social, political, and ethical challenges associated with medical AI, such as issues of trust, privacy, explainability, bias, equity, and responsibility in relation to AI technologies.

**Keywords:** Embedded Ethics · Artificial Intelligence · Medical Imaging · Human-Computer-Interaction

## 1 Introduction

### 1.1 AI in Medicine

We have long known about the potential of so-called *artificial intelligence* (AI), but with the release of *ChatGPT*, a chatbot or text-based dialog system from the U.S. company *OpenAI*, in November 2022, the world has once again witnessed the capabilities these modern technologies hold. Thanks to advances in *machine learning* (ML) and *deep learning* (DL), AI can now process, analyze, and interpret data in very short time and thereby preparing and enhancing human decision-making.

Especially in the medical context, AI is seen as a *key technology*: Computer programs, for example, learn to predict the individual course of illness and therapy by means of AI-supported analysis of an infinite number of medical records. Intelligent assistance systems and care robots assist people with limited mobility. Medical wearables measure, record and interpret the patient’s vital signs contributing to the continuous monitoring of chronic diseases. Hence, AI-based prognostics and methods not only help with *diagnosis* and *therapy*, but also with the reliable and yet cost-efficient *care* of patients (Siontis et al. 2021).

AI is ubiquitous. It is making its way into our everyday practices. By doing so, it is not only changing the way we *perceive* and *interact* with our environment but is *transforming* it at the same time. This becomes particularly evident in the medical context (cf. Rajpurkar et al. 2022): AI in medicine is not only changing the way of patient examination, perception of that examination and the interaction with physicians, it is also reshaping the very context in which we live. By using AI-based remote diagnostic tools or optimizing the processes of existing medical infrastructure, medical care can be improved in large parts of the world and the overall human right to health can be protected more consistently (Raso et al. 2018).

Although the use of AI in medicine may sound promising at first, ethical considerations point to certain risks of the current use and design of AI in medicine: Lack of transparency, explanation, and fairness, but also insufficient protection of patients' privacy and their sensitive health data are just a few examples of the specific challenges in dealing with medical AI.

## 1.2 Objective: Interdisciplinary Research on the Social and Ethical Aspects of Medical AI

As AI continues to advance and permeate various aspects of our lives, it is crucial that we consider the ethical and social implications of these developments. From issues of bias and discrimination, privacy and autonomy, transparency, and accountability, to questions of human-machine-interaction, the ethical and social issues surrounding AI are complex and multifaceted and need to be addressed carefully and responsibly.

Due to the interdisciplinary character of AI itself and its application contexts (cf. Zhuang et al. 2020), we assume that AI and its impact can only be comprehensively researched on an *interdisciplinary* basis. Accordingly, AI must be problematized and analyzed in light of various disciplines—such as *computer science*, *science and technology studies*, *medicine*, and *philosophy*. By exploring the ethical and social aspects of AI in such cross-disciplinary frameworks, it is possible to draw on the different expertise and findings without disregarding the respective focus, approaches, and methods of each discipline.

The overall objective of this project can be classified on three levels:

- (*Descriptive level*) Identifying conflicts of interest between stakeholders within health-care AI innovation (e.g., physicians, patients, health insurers, care-givers).
- (*Theoretical level*) Uncovering and deconstructing key terms and philosophical concepts—such as autonomy, vulnerability, explainability or responsibility—as well as their mutual relation within sociotechnically transformed practices.
- (*Normative level*) Developing concrete proposals for the responsible and trustworthy use and design of AI in the medical context based on *empirical* and *hermeneutic research*.

## 2 ‘CReAIte’ and Its Pilot Project ‘MedAIcine’

### 2.1 ‘CReAIte’: Center for Responsible AI Technologies

The Center for Responsible AI Technologies (*CReAIte*), founded by the Technical University of Munich (TUM), the Munich School of Philosophy and the University of Augsburg in February 2022, pursues the goal of incorporating philosophical, ethical, and social science inquires throughout the process of developing, implementing, and critically reflecting on AI technologies. *CReAIte* has set itself the task of rethinking human-machine interaction and contributing to a better understanding of the transformative power of technologies, such as ML and DL.

*CReAIte* intends on expanding this integrated view on interdisciplinary AI research and its (techno)philosophical, ethical as well as societal and political dimensions to four different—albeit crucial—fields of application: (i) Medicine/Care/Health; (ii) Future of work; (iii) Mobility; (iv) Climate/Environment.

By involving politics, ethics, law, computer science, science and technology studies as well as cultural studies, *CReAIte* aims to facilitate the emergence of technical innovations that can perform the tasks assigned to them *reliably*, but also in a *socially responsive* and *responsible* manner.

### 2.2 ‘MedAIcine’: Pilot Project

With the research on the social and ethical aspects of AI in medicine, *CReAIte* started its pilot project: *MedAIcine*. In this particular project the main focus is on the use of AI in medical imaging, such as X-ray, MRI or CT.

Especially in the field of imaging diagnostics, AI systems are already widely used and researched. According to preliminary results, integrating AI and *computed-aided detection* (CAD) with screening methods, in fact, has shown reliable and accurate screening results (Goyal et al. 2020, 18). However, *MedAIcine* also highlights the aforementioned conflicts and challenges of embedding AI technologies into our practices: For example, if AI systems are trained with data sets that do not represent all skin colours and ethnicities, patterns such as tumour thickness or the size of a suspected melanoma cannot be recognised and determined equally well for all skin colours and ethnicities either. As a result, patients are being discriminated, putting some people at a disadvantage when it comes to medical care.

## 3 Use Cases and Methods

### 3.1 Embedded Ethics

In *MedAIcine*, we apply an innovative interdisciplinary approach known as embedded ethics and social science (Breuer et al. 2023). ‘Embedded ethics’ denotes a research practice that involves an ongoing integration of ethical and social analyses into the entire development process (McLennan et al. 2020). As a research team comprising scholars from science and technology studies (STS), philosophy, and ethics, we share a common interest in investigating the complex social and ethical issues arising in the

development of machine learning systems in medical imaging. We study these issues empirically, leveraging long-term integrated collaboration with engineering researchers and medical practitioners.

We use a qualitative, inductive, and interpretivist approach, following grounded theory methodology (Charmaz 2006) with an iterative process of data collection and analysis. For data collection, we conduct ethnographic field studies where we write field notes as embedded, overt, participant observers at AI research labs and in hospitals; we obtain pseudonymized qualitative semi-structured interviews with AI researchers, medical experts, and patients, as well as scenario-based focus groups with patients. Throughout the process, we adapt our sampling strategy based on insights from ongoing analysis of our incoming data. In our analysis, we apply analytical lenses from STS, philosophy, and ethics to come to a rich understanding of the practices at hand in the development and clinical implementation of medical AI.

Ultimately, we aim for interdisciplinary co-design of medical AI applications, where ethical and social analyses constitute integral parts of design processes and workplace integration. This empirically based approach thus constitutes an alternative to the prevalent more abstract, planning- and, principle-oriented efforts in technology ethics and innovation that have often fallen short of expectations to ensure responsible conduct in research and development (Winfield and Jirotko 2018).

In its empirical orientation, our approach draws on and complements traditions of parallel and collaborative interdisciplinary research. It resonates with approaches of value-sensitive design (Friedman et al. 2012) and relational empirical ethics of care technologies (Pols 2015). It shares with ethics parallel research (van der Burg 2009) the interest in investigating social and ethical aspects alongside and in concert with research in science and engineering. Yet, embedded ethics and social science goes beyond this goal in its aim to achieve long-term *integration* of social, ethics, science, and technology research. Our approach is thus squarely rooted in a tradition of sociotechnical integration research (Fisher and Schuurbiers 2013; Fisher et al. 2015). It draws from the field of STS viewing emerging technologies as complex, sociotechnical compounds that mediate interactions between researchers, developers, users, and other affected or involved stakeholders and machines-interactions we aim to better understand by looking at concrete case studies.

### 3.2 Use Cases: Radiology, Endoscopy, and Dermatology

**Radiology: AIM Lab.** The first use case we focus on is ML in radiology. We investigate this field by way of a case study of a university research laboratory, the Lab for Artificial Intelligence in Medicine (*AIMLab*) at TUM. Into the *AIMLab*, embedded STS researchers are integrated, conducting a social science laboratory field study. They undertake regular lab visits, attend lab meetings, shadow AI researchers in their everyday work and conduct peer-to-peer interviews with them. The combination of computer science research tradition and an university hospital radiology department allows us valuable insights into the synergies and tensions that emerge between disciplinary research and an application domain. Focus of this case study is to gain an in-depth understanding of

the particularities of the *practices* in AI research for radiology, as well as the specific social and ethical issues that arise in relation to this domain.

**Endoscopy: Achalasia.** The second use case being researched in *MedAIcine* addresses the diagnosis and treatment of achalasia—a rare dysfunction of the esophageal muscles and the lower sphincter. In the gastroenterology department of Augsburg University Hospital, physicians are leading a study concerning the 3D-reconstruction of the esophagus with a multimodal data system. Their goal is to program a model of the esophagus using specially developed algorithms, to specify its stretching ability, inclination, and muscle thickness. In accordance with the embedded ethics approach described above, the researchers involved in *MedAIcine* are already engaged in the early phase of the development of such a prototype for the digital, AI-supported augmentation of human organs. Thus, ethical research on possible standards of fair training data or a sufficient degree of explainability of technology is involved even before the AI-systems are programmed.

**Dermatology: OCTOLAB.** Further, we are embedded into the OCTOLAB project of the University Hospital Augsburg, where optical coherence tomography for the diagnosis of basal cell carcinomas is to be integrated into a long-pulsed infrared laser for the therapy of basal cell carcinomas. The associated diagnostics and therapy will be based on AI. The aim of the project is that the combined device will contribute to automated diagnostics and therapy in the early detection and individualized minimally invasive therapy of basal cell carcinomas. Our focus of this case study is to gain a deeper understanding of the use of the AI system in (changing) medical practice, exploring in particular ethical implications of using a *closed-loop system* for automated diagnostics and therapy. To achieve this, we participate in consultations with patients and meetings of medical and scientific stakeholders involved in OCTOLAB. Additionally, we conduct *scenario-based focus group discussions* with physicians and patients.

## 4 Preliminary Findings and Emerging Topics

As a reaction to socio-technical changes in medical contexts and practices, the preliminary findings of the interview study (*peer-to-peer researcher interviews*) at *AIMLab* already show that people often demand the respective technology to be explainable. Especially in critical areas of application, such as military defense or medicine, transparency regarding the reasoning of the AI-based decision-making process are seen as essential. The output of an AI-system should therefore be visible as well as comprehensible to other external agents. For instance, the final decision, such as a recommendation for emergency surgery or a patient's cancer diagnosis, should remain comprehensible and understandable to medical staff. *Explainability* is therefore stressed as a key aspect for developing transparent and trustworthy AI, leading to a thriving stream of research: the so-called 'eXplainable Artificial Intelligence' (XAI). It is assumed that sufficient explainability may overcome the black box issues of opaque AI systems. But when is something considered sufficiently justified or comprehensible? Which criteria constitute explainability, as it is often emphasized in the context of AI? The interviews conducted so far indicate that, even among experts, no shared understanding of (sufficiently) explainable AI prevails yet. Gaining insights into a variety of explainability and interpretability

techniques employed in the *AIMLab*, will help us investigate the concrete research and technical practices related to making ML-machine learning models interpretable and explainable. Complementing our empirical research, philosophical reflections on *explanation*—as raised in the history of ideas or the philosophy of science—here can contribute to provide a conceptual basis for the discussion on medical XAI. In philosophy different schemes of explanation are distinguished (e.g., *statistical relevance-model* or *causal mechanical-model*). In a generic sense, however, all philosophical concepts think of explanations as a *linguistic and logical construct*, which reveals the central causes of a certain phenomenon and thereby demonstrates its *causality* (Hocutt 1974, 385). By means of formal logical reconstructions, a kind of *regularity* (if  $p$ , then  $q$ ) is to be made explicit, which establishes a logical (causal) relation between cause  $p$  (*explanans*) and effect  $q$  (*explanandum*) (Ruben 2004, 110).

In the further course, *MedAlcine* will explore the limitations of these explanatory models—hitherto known in philosophy—in the context of modern technologies, such as AI, and strives for refining a new understanding of explanatory power in contrast to correlated yet distinct concepts, such as reliability, transparency or understanding (cf. Abel 1948).

Moreover, issues concerning the altering relationship between humans and machines in general and the human-computer-interaction in particular emerged in the course of the interviews. For instance, how is medical AI affecting established—in the medical context, mostly asymmetrical—power and trust relationships? Against this backdrop, *MedAlcine* is currently exploring these possible transformations along considerations of *power* and ethics on *vulnerability*. Being and becoming vulnerable by technology can generally be understood on three levels: Due to its very essence all humans are vulnerable (*ontological vulnerability*)—some moreover by situation (*situational vulnerability*) or by structure (*structural vulnerability*). Since all people are to be understood as vulnerable, the focus of thinking is less on risk and more on the positive direction of understanding people as vulnerable beings (cf. Haker 2021). By recognising the ontological vulnerability of all individuals—in the sense of a *conditio humana*—and the situational and structural factors that can exacerbate vulnerability, we can approach the use of AI systems in medicine with empathy and a commitment to protecting the welfare of all involved (i.e., patients, medical staff, companies producing medical technologies etc.).

In addition to making people vulnerable via technology, *MedAlcine* examines power structures in the context of medical human-machine interactions. Drawing on emerging theories such as data colonialism (cf. Ricaurte 2019) and data feminism (D’Ignazio and Klein 2020), *MedAlcine* explores power structures underlying the development, use and design of medical AI. Here, scholars in these fields emphasize to reflect on the initial mechanisms and the extent to which these power structures are altered and (re)produced through algorithmic biases.

## 5 Concluding Thoughts

In our project, *MedAlcine*, we investigate various aspects that pertain to Human-Computer-Interaction (HCI). In exploring human-centered AI design, our focus lies on studying philosophical and ethical implications of the HCI regarding specialized AI

systems in imaging diagnostics. Grounded on our ‘embedded ethics and social science’ research, we offer an interdisciplinary, empirically, and philosophically informed sensitivity to issues of explanation, trust, and power relations arising in various dimensions of HCI. We are convinced that reflections of the philosophy of AI and STS analyses on behavioral change in context of AI contribute to advance our understanding of HCI. We are eager for an exchange on these matters with the community at HCI International 2023.

## References

- Abel, T.: The operation called ‘Verstehen.’ *Am. J. Sociol.* **54**(3), 211–218 (1948)
- Breuer, S., Braun, M., Tigard, D., Buyx, A., Müller, R.: How engineers’ imaginaries of healthcare shape design and user engagement: a case study of a robotics initiative for geriatric healthcare AI applications. *Trans. Hum.-Comput. Interact. (TOCHI)* **30**(2), 1–33 (2023). <https://doi.org/10.1145/3577010>
- Charmaz, K.: *Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis*. SAGE Publication, London, Thousand Oaks, New Delhi, Singapore (2006)
- D’Ignazio, C., Klein, L.F.: *Data Feminism*. Strong Ideas Series. The MIT Press, Cambridge (2020)
- Fisher, E., O’Rourke, M., Evans, R., Kennedy, E.B., Gorman, M.E., Seager, T.P.: Mapping the integrative field: taking stock of socio-technical collaborations. *J. Responsible Innov.* **2**(1), 39–61 (2015). <https://doi.org/10.1080/23299460.2014.1001671>
- Fisher, E., Schuurbijs, D.: Socio-technical integration research: collaborative inquiry at the mid-stream of research and development. In: Doorn, N., Schuurbijs, D., van de Poel, I., Gorman, M.E. (eds.) *Early Engagement and New Technologies: Opening up the Laboratory*. PET, vol. 16, pp. 97–110. Springer, Dordrecht (2013). [https://doi.org/10.1007/978-94-007-7844-3\\_5](https://doi.org/10.1007/978-94-007-7844-3_5)
- Friedman, B., Kahn, P., Borning, A.: *Value sensitive design: theory and methods*. University of Washington Technical report, 02-12 (2002)
- Goyal, H., et al.: Scope of artificial intelligence in screening and diagnosis of colorectal cancer. *J. Clin. Med.* **9**(10), 3313 (2020)
- Haker, H.: Verletzliche Freiheit. Zu einem neuen Prinzip der Bioethik. In: Keul, H. (eds.) *Theologische Vulnerabilitätsforschung. Gesellschaftsrelevant und interdisziplinär*, pp. 99–118. Kohlhammer, Stuttgart (2021)
- Hocutt, M.: Aristotle’s four because. *Philosophy* **49**(190), 385–399 (1974)
- McLennan, S., et al.: An embedded ethics approach for AI development. *Nat Mach Intell* **2**(9), 488–490 (2020). <https://doi.org/10.1038/s42256-020-0214-1>
- Pols, J.: Towards an empirical ethics in care: Relations with technologies in health care. *Med. Health Care Philos.* **18**(1), 81–90 (2015)
- Raso, F.A., Hilligoss, H., Krishnamurthy, V., Bavitz, C., Kim, L.: *Artificial intelligence & human rights: opportunities & risks*. Berkman Klein Center Research Publication (2018)
- Rajpurkar, P., Chen, E., Banerjee, O., et al.: AI in health and medicine. *Nat. Med.* **28**, 31–38 (2022). <https://doi.org/10.1038/s41591-021-01614-0>
- Ricarte, P.: Data epistemologies, the coloniality of power, and resistance. *Telev. New Media* **20**(4), 350–365 (2019). <https://doi.org/10.1177/1527476419831640>
- Rubén, D.: *Explaining Explanation*. Routledge, London/New York (2004)
- Siontis, K.C., Noseworthy, P.A., Attia, Z.I., et al.: Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nat. Rev. Cardiol.* **18**, 465–478 (2021). <https://doi.org/10.1038/s41569-020-00503-2>

- Winfield, A.F.T., Jirotko, M.: Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philos. Trans. Ser. A Math. Phys. Eng. Sci.* **376**(2133) (2018). <https://doi.org/10.1098/rsta.2018.0085>
- van der Burg, S.: Imagining the future of photoacoustic mammography. *Sci. Eng. Ethics* **15**(1), 97–110 (2009). <https://doi.org/10.1007/s11948-008-9079-0>
- Zhuang, Y., Cai, M., Li, X., Luo, X., Yang, Q., Wu, F.: The next breakthroughs of artificial intelligence: the interdisciplinary nature of AI. *Engineering* **6**(3), 245–247 (2020)