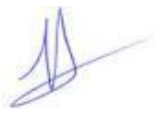
A European Cancer Image Platform Linked to Biological and Health Data for Next-Generation Artificial Intelligence and Precision Medicine in Oncology

Deliverable D1.5:

Guidelines of ethics and social implications of AI in oncologic imaging

Reference	D1.5_ EuCanImage_BBMRI-ERIC_28092023
Lead Beneficiary	BBMRI-ERIC
Author(s)	Melanie Goisauf & Mónica Cano Abadía
Dissemination level	Public
Type	Report
Official Delivery Date	28/09/2023
Date of validation of the WP leader	28/09/2023
Date of validation by the Project Coordinator	28/09/2023
Project Coordinator Signature	

EuCanImage is funded by the European Union's H2020 Framework Under Grant Agreement No 952103

Version log

Issue Date	Version	Involved	Comments
08/09/2023	1	Melanie Goisauf, Mónica Cano	First draft
27/09/2023	2	Melanie Goisauf, Mónica Cano, Oliver Díaz, Xènia Puig Bosch, Michaela Th. Mayrhofer	Final draft
28/09/2023	Final	Xènia Puig Bosch, Oliver Díaz, Karim Lekadir	Revised and corrected final version

Executive Summary

Artificial intelligence (AI) is being applied in the medical field to improve healthcare, but it also poses challenges, particularly in terms of ethics and societal implications. It raises complex questions surrounding informed consent, biases leading to inequality, data privacy and protection risks, as well as responsibility and liability concerns. To address these issues, several guiding principles and recommendations have been formulated based on key ethical values. However, these principles need to be adapted and applied specifically to each field of application – in the case of EuCanImage AI systems in radiology and oncology. This deliverable presents the results of an analysis conducted on the ethical and societal implications of AI in oncologic imaging. The analysis utilizes empirical data gathered in WP1 of EuCanImage and focusses on trustworthy AI as a key concept. Aligned with the FUTURE-AI initiative, this deliverable provides guidelines and tools for AI systems design, including considerations of potential biases (especially regarding sex and gender dimensions), an interdisciplinary and embedded ethics approach, the importance of considering situated practices, stakeholder engagement, as well as environmental aspects.

Acronyms

Name	Abbreviation
Artificial Intelligence	AI
Artificial Intelligence High Level Expert Group	AI - HLEG
Artificial Intelligence in Medicine Lab	BCN-AIM
Biobanking and Biomolecular Resources Research Infrastructure – European Research Infrastructure Consortium	BBMRI-ERIC
European Commission	EC
European General Data Protection Regulation	GDPR
Machine Learning	ML
Organisation for Economic Co-operation and Development	OECD
Work package	WP
World Health Organization	WHO

Table of Contents

List of Tables

1 Purpose of the deliverable

The overall objective of WP1 in EuCanImage is to develop a governance framework that considers ethical issues, legal requirements and societal aspects of transnational data sharing as well as for the development and utilization of AI-supported image-based decision support tools in clinical oncology. This deliverable focuses on the ethical and social implications especially. In particular, it presents the analysis of ethical and social implications of AI-based cancer imaging solutions and contributes to the WP goals in translating the outcomes into guidelines for AI systems design.

2 Introduction

AI systems will have a significant impact in the medical field as they are expected to improve diagnostic performance. Besides technical challenges, the application of these systems raises questions regarding the ethical and societal implications.

The potential far-reaching consequences of AI in several areas of society are addressed in the evolving field of 'Ethics of AI' (see Dubber, Pasquale, & Das, 2020). There are discussions about possible harmful consequences of AI use for individuals and groups, especially for the most vulnerable populations, and on the risk of perpetuating or even amplifying ethical and societal injustices. For medicine and healthcare, and radiology more specifically, this raises complex questions concerning the scope of informed consent, biases that may result in inequality, and risks associated with data privacy and protection, as well as open questions regarding responsibility and liability.

Based on key ethical values such as respect, autonomy, beneficence, and justice (Beauchamp & Childress, 2001), several guiding principles and recommendations have been formulated to tackle these issues (Currie, Hawk, & Rohren, 2020; Ryan & Stahl, 2020). Such principles and recommendations have also been communicated on EU level (High-Level Expert Group on Artificial Intelligence, 2019) or in initiatives such as FUTURE-AI[1] (Lekadir et al., 2021).

High-level principles need to be adapted and situated in a specific field of application – in the case of EuCanImage, AI systems in radiology and oncology – and translated into practice-oriented guidelines. In the scientific discourse on ethics of AI (see Goisauf & Cano Abadía, 2022) it becomes clear that the interplay of the technology and social frameworks, systemic structures and power relations that intersect with identity traits (e.g., gender, race, socio-economic status), as well as the implications of private ownership and the role of corporations, profit-making, and geopolitical structures need to be considered in a responsible development of medical AI.

This document outlines the results of an analysis conducted on the ethical and societal implications of AI in oncologic imaging, utilizing empirical data gathered from the EuCanImage project. Against the background of a plethora of implications that have been identified in the ethics of AI discourse so far, this deliverable focuses on requirements concerning trustworthy

[1] See: https://future-ai.eu/ (accessed on 22.08.2023)

AI as one of the key concepts. The considerations outlined in this document are closely connected to FUTURE-AI, in building on and elaborating central ethical and social aspects.

3 Methodology

To investigate the ethical and social implications of AI in radiology, we designed and conducted a broad empirical study in which rich qualitative data were collected and analyzed. This research design of this study received ethics approval by the Ethics Committee of the Medical University of Graz (EK-Number 33-650 ex 20/21).

The empirical research consists of:

Systematic review of state-of-the art academic literature on ethics of AI in radiology

We performed a comprehensive review of ethical and societal issues that have already been identified and discussed, as well as on how these issues have been addressed in the context of AI. We carried out a systematic review of state-of-the art academic literature between July and December 2021. Five search engines were used (Google Scholar, Microsoft Academic, PubMed, Scopus, and Web of Science) to identify relevant articles on these issues.

After screening the relevant records, the full texts of the resulting sample (n=56) were analyzed using thematic analysis (Terry, Hayfield, Clarke, & Braun, 2017). Each article in the final sample was open coded to identify overarching themes and patterns and analyzed to determine their specific content and depth, but also conceptual gaps. The findings of the review have been published open access in Frontiers in Big Data.

Qualitative interviews

Thirteen qualitative interviews were conducted with experts of the EuCanImage consortium and beyond. Eleven interviews were conducted face-to-face during a research stay at the Artificial Intelligence in Medicine Lab (BCN-AIM) at the University of Barcelona, and two interviews were conducted via Zoom.

The interviews were conducted based on a topic guide that covered work practices and areas of expertise as well as attitudes towards the role, benefits and limits of radiological AI. Interview partners were AI and platform developers, clinicians, and patient representatives. The transcribed interviews were analyzed in-depth using the Atlas.ti software and based on established qualitative methods, in particular open coding and analytic approaches from Grounded Theory Methodology (Charmaz, 2006).

Expert workshop on the ethical and social implications of AI in biomedical research on cancer

In building on the findings of the literature review, we organized an expert workshop on the ethical and social implications of AI in biomedical research on cancer, 5-6 May 2022 in Berlin, Germany. The workshop used the synergies of three projects, EuCanImage and INTERVENE (Horizon 2020; Grant agreement ID: 101016775), as well as Bigpicture (IMI; Grant agreement ID: 945358.). Experts from these projects and beyond discussed topics such embedding ethics into the development of AI, algorithmic impact assessment, predicting future diagnosis (Electronic Health Records), biases in training and calibration of AI, explainability and trustworthiness of AI solutions in cancer imaging and polygenic risk score generation and use, as well as clinical application. Key discussion points were collected on a digital mind-map following the ECOUTER stakeholder engagement methodology (Murtagh et al., 2017). We identified trustworthiness as a principle that needs further attention and invited the experts to fill in an open questionnaire to deepen insights into key requirements, challenges, and conditions on the matter. The results of this co-creative exercise will be published in a paper.

EuCanImage consortium workshops

EuCanImage consortium members participated in different workshops on trustworthiness and bias during:

- The two-year project meeting in Barcelona in October 2022;
- An inter-work package meeting online in November 2022;
- And the M30 consortium meeting in Vienna in January 2023.

Key discussion points were collected through mapping (on-site and online) following the ECOUTER stakeholder engagement methodology (Murtagh et al., 2017). The findings and expertise gained in these research activities are incorporated into the development of the FUTURE-AI principles. Researchers from BBMRI-ERIC working in WP1 are involved in the FUTURE-AI initiative.

4 Ethical and social/societal implications of AI in oncologic imaging

4.1 Scientific discourse on ethical and social/societal implications of medical AI

The review of the scientific discourse on radiological AI (Goisauf & Cano Abadía, 2022), which we summarize and reflect critically from an ethics and social science perspective in the following paragraphs, shows that the key topics discussed are about their potential to improve predictive analytics, diagnostic performance, and eventually patient outcomes, as well as challenges that arise due to the (potential) real-world application. Focusing on ethical and societal implications described in the literature, major themes are expectations and challenges regarding the application of AI systems in the medical field, and related ethical principles such

as explainability, interpretability, trust/trustworthiness, responsibility and accountability, justice, and fairness.

Ethical and social/societal implications are reflected in the benefits described for end users. Therefore, it is expected that AI systems in healthcare will improve the health(care) of populations, reduce the cost of healthcare, and improve the work life of healthcare providers.

Nonetheless, the socio-technological conditions under which these expectations can be met, and, at the same time, challenges (such as bias and black-box) can be managed are unclear. We identified a need for more interdisciplinary research on bias in radiology and a deeper investigation of several ethical and societal implications of AI use to avoid potential discriminatory effects beyond and as part of technical solutions. The findings of the review have been published open-access in Frontiers in Big Data. Two major challenges were identified in the review: black box and bias.

Black box

A black box is understood as "an apparatus whose inner-workings remain opaque to the outside observer" (Quinn, Jacobs, Senadeera, Le, & Coghlan, 2021, p. 2). Opacity, intelligible justifications, and recommendations are key issues for medical AI that need to be discussed when considering ethical requirements and the practitioner-patient relationship. Ferretti et al. (2018) frame the problem of black boxes in medicine by applying the concept of opacity, which can be differentiated into three types: (1) lack of disclosure, (2) epistemic opacity, and (3) explanatory opacity. To ensure the ethical use of data and to address a lack of disclosure, "patients should know who has access to their data and whether (and to what degree) their data has been deidentified. From an ethical perspective, a patient should be aware of the potential for their data to be used for financial benefit to others and whether potential changes in legislation increase data vulnerability in the future, especially if there is any risk that the data could be used in a way that is harmful to the patient" (Currie et al., 2020, p. 749). Epistemic consequences of black-box medicine may be a loss of knowledge, and specifically to a loss of medical understanding and explanation and, thus, medical advances.

An opaque system makes it difficult to keep humans in the loop and enables them to detect errors and to identify biases. Such a system can have negative effects on underrepresented or marginalized groups and can also fail in clinical settings (Quinn et al., 2021). In addition, it can pose certain risks for radiologists, who are expected to validate something that they cannot understand (Neri, Coppola, Miele, Bibbolino, & Grassi, 2020), open them to adversarial attacks (Geis et al., 2019; Tizhoosh & Pantanowitz, 2018), or intensify the clash between black-box medicine and the duty of care, presuming that the radiologists have the ability to understand the technology, its benefits, and potential risks. The latter is also associated with depriving the patients of the ability to make decisions based on sufficient information and justifications, which contradicts the ethical requirement for the patients to exercise autonomy by giving their informed consent (Quinn et al., 2021).

Bias

Besides technical biases, AI systems used in healthcare might have both a racial and a gender bias (Rasheed et al., 2022). Many algorithms in medicine have been shown to encode,

reinforce, and even exacerbate inequalities within the healthcare system (Owens & Walker, 2020) and can worsen the outcomes for vulnerable patients (Quinn et al., 2021). Such biases are introduced due to the data used to train an algorithm and the labels given to these data, which may be laden with human values, preferences, and beliefs (Geis et al., 2019). The generated outputs will thus eventually reflect social and political structures, including injustices and inequalities. Consequently, AI systems cannot provide entirely unbiased or objective outcomes based on incomplete or unrepresentative data; instead, they mirror the implicit human biases in decision-making (Abràmoff, Tobey, & Char, 2020; Balthazar, Harri, Prater, & Safdar, 2018; Pesapane, Volonté, Codari, & Sardanelli, 2018; Ware, 2018). This has effects that extend beyond training, an aspect underlined by Quinn et al. (2021, p. 4), who point out that "most training data are imperfect because learning is done with the data one has, not the sufficiently representative, rich, and accurately labeled data one wants. [...] even a theoretically fair model can be biased in practice due to how it interacts with the larger healthcare system."

Common sources of bias that potentially promote or harm group level subsets are based on gender, sexual orientation, ethnic, social, environmental, or economic factors, as well as on unequal access to healthcare facilities and geographical bias. One solution described in the literature is to ensure diversity when collecting data and to address bias in the design, validation, and deployment of AI systems. Furthermore, careful performance analyses should be performed on the basis of population subgroups, including age, ethnicity, sex, sociodemographic stratum, and location.

Understanding the impact of a new algorithm is particularly important; this means that, if the disease spectrum detected using the AI system differs from that identified using current clinical practice, then the benefits and harms of detecting this different disease spectrum must be evaluated (Kelly, Karthikesalingam, Suleyman, Corrado, & King, 2019, pp. 4-5). Geis et al. (2019, p. 331) propose certain questions that can be asked to identify bias to advance toward algorithmic fairness: How and by whom are labels generated? What kinds of bias may exist in the datasets? What are the possible risks that might arise from those biases? What steps have we taken to mitigate these risks?

A critical reflection of the reviewed literature shows that bias has not been framed in the context of power relations and societal conditions, nor has it been referenced to the existing body of research on, e.g., how gender and race shapes and affects biomedicine and healthcare practice (Kaufman, 2013; Oertelt-Prigione & Regitz-Zagrosek, 2012; Roberts, 2008; Schiebinger & Schraudner, 2011) or how gender and racial bias in algorithms could have a negative impact in certain areas of society (e.g. Noble, 2018; O'Neil, 2016). Bias has been shown to affect every stage of data processing (i.e., in generating, collecting, and labeling data that are used to train AI tools) and to affect the variables and rules used by the algorithms. Hence, AI tools can be taught to discriminate, reproduce social stereotypes, and underperform in minority groups, an especially risky proposition in the context of healthcare (Char, Shah, & Magnus, 2018; Wiens et al., 2019).

Researchers are increasingly recognizing the importance of addressing bias in datasets by promoting diversity (Leavy, 2018). However, simply ensuring diversity is not sufficient (Li et al., 2022). Further research is needed to understand how discrimination and socioeconomic factors intersect, as they can introduce bias into healthcare algorithms through societal inequalities (Quinn et al., 2021). It is considered best practice to anticipate structural bias in datasets and comprehend the social implications of using AI systems before implementing

them. Some authors (Owens & Walker, 2020) even argue that failing to do so should be classified as scientific misconduct.

4.2 Guiding principles in the scientific discourse

During the analysis of the scientific discourse, certain topics were identified as especially important, which are mainly organized around approaches and principles. Guided by our research questions (i.e., what types of ethical issues are raised by medical AI and how these are tackled in radiology and the case of breast cancer in particular), we analyzed the key themes regarding their claims about ethical and societal implications.

Explainability and interpretability

To manage the risks inherent in the use of medical black boxes and the resulting bias, the requirement is often posed that the way an AI system arrives at its decision must be transparent and sufficiently understandable for the "human-in-the-loop" to improve patient safety and to gain the patient's trust. For that reason, "explainability" has become a key principle in AI ethics, and especially in the context of healthcare.

Interpretability refers to how well one can understand how an AI system works, while explainability refers to how well one can explain what happens in AI decision-making in understandable terms (Brady & Neri, 2020; Rasheed et al., 2022).

The expectation for AI should be that "AI can explain itself at least as well as human explain their own actions and reasonings, systems would demonstrate transparency and honesty" (Ware, 2018). The European General Data Protection Regulation (GDPR) has emphasized the patient's right to receive an explanation as a top priority in ML research. The right to an explanation encompasses the right to receive an explanation about the outputs of the algorithm, especially when decisions need to be made that significantly affect an individual.

Trust and trustworthiness

Many guidelines, including the FUTURE-AI guidelines, point out trust and trustworthiness as one of the key principles. As such, it has been chosen as one of the guiding principles by the High-Level Expert Group on AI (AI HLEG) of the EC and identified as the defining paradigm for their ethics guidelines (High-Level Expert Group on AI, 2019).[2]

Trustworthiness as a concept appears for the AI HLEG as something that is achieved when certain bioethical principles and AI particularities are followed. In this sense, an Assessment

[2] See the two AI HLEG guidelines produced in 2019. Ethics guidelines for trustworthy AI: https://digitalstrategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (accessed on 23.08.2023) and Policy and Investment Recommendations for Trustworthy AI: https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_policy_and_investment_recommendations.pdf (accessed on 23.08.2023)

List for Trustworthy AI[3] has been created to help assess whether the AI system complies with the seven requirements of trustworthy AI:

- human agency and oversight;
- technical robustness and safety;
- privacy and data governance;
- transparency;
- diversity, non-discrimination, and fairness;
- environmental and societal well-being;
- and accountability.[4]

The AI HLEG mirrors the bioethical principles proposed by Beauchamp and Childress (2001) by identifying the principles of beneficence (doing good and no harm), autonomy (preserving human agency), and justice (being fair[5]).

The WHO has published ethical guidance on the use of AI for health, where they highlight the importance of the engagement and the role of the public, as "the effective use of AI for health will require building the trust of the public, providers, and patients" (World Health Organization, 2021, p. 69). Similarly, the OECD launched a policy observatory in 2020 that "aims to help countries enable, nurture and monitor the responsible development of trustworthy artificial intelligence systems for the benefit of society".[6] The organization has developed the G20 AI Principles that are value-based and aim at fostering innovation and trust in AI. They identify five complementary value-based principles: inclusive growth, sustainable development and well-being; human-centered values and fairness; transparency and explainability; robustness, security and safety; and accountability.[7] In line with the G20 AI Principles, frameworks for health data governance are expected to emphasize transparency, public communication and stakeholder engagement, explicitly highlighting the importance of trust.[8] In this regard, according to the OECD, "lack of trust among patients, the public, data custodians and other stakeholders, in how data are used and protected is a major impediment to data use and sharing".[9]

The reviewed literature shows that there is certain level of consensus concerning the fact that black boxes and the lack of interpretability and explainability can lead to a lack of trust(worthiness) in and acceptance of AI systems by clinicians and patients. Consequently, AI systems should be transparent enough that those using them can have access to the processes that govern them and be able to explain them. This requires access to accessible, intelligible, and usable information that can be effectively evaluated. In turn, a lack of explainability, lack of transparency, and lack of human understanding of how AI systems work will inevitably result in clinicians failing to trust decisions made by AI, as well as failing to

[3] https://altai.insight-centre.org/ (accessed on 23.08.2023)
[4] See footnote 2 for AI HLEG documents.
[5] Fair, i.e. being fair, in contrast to FAIR (Wilkinson et al., 2016) which focuses on the machine-actionability of data being findable, accessible, interoperable and reusable.
[6] See page 331 of the OECD's AI Policy Observatory fact sheet: https://www.oecd.org/going-digital/ai/about-the-oecd-ai-policyobservatory.pdf (accessed on 23.08.2023).
[7] See page 15 of the OECD's Trustworthy AI in health: Background paper for the G20 AI dialogue, digital economy, and trade: https://www.oecd.org/health/trustworthy-artificial-intelligence-in-health.pdf (accessed on 23.08.2023).
[8] See OECD's Recommendation of the Council on Health Data Governance, OECD/LEGAL/0433: http://legalinstruments.oecd.org (accessed on 23.08.2023)
[9] See page 16 of the OECD's Trustworthy AI in health: Background paper for the G20 AI dialogue, digital economy, and trade: https://www.oecd.org/health/trustworthy-artificial-intelligence-in-health.pdf (accessed on 28.08.2023)

trust the reliability and accuracy of such systems (Bjerring & Busch, 2021; Larasati & DeLiddo, 2020). Specific challenges have been identified as decreasing trust in AI in medicine: patient harm due to AI errors, misuse of medical AI tools, risk of bias in medical AI and perpetuation of inequities, lack of transparency, privacy and security issues, gaps in AI accountability, obstacles to implementation in real-world healthcare (Lekadir, Quaglio, Tselioudis Garmendia, & Gallin, 2022).(Bjerring & Busch, 2021; Larasati & DeLiddo, 2020).

Trust is not simply achievable by balancing out certain technical features. Trust can be understood "a fundamental principle for interpersonal interactions and [...] a foundational precept for society to function" (Ryan & Stahl, 2020, p. 74). In this sense, trust is a complex, situated, context-dependent, and relational concept that involves several trustor/trustee relationships, such as trust in persons (e.g., scientists who trust each other, patients who trust scientists and clinicians), technology, and institutions (Bijker, Sauerwein, & Bijker, 2016; Wyatt, Harris, Adams, & Kelly, 2013). Trust refers to being open to vulnerability due to having positive beliefs about someone's intentions or actions (Baier, 1986). Trust simplifies the process of decision-making by streamlining the gathering and analysis of information. Additionally, trust influences behavior by indicating the most practical and advantageous actions based on the assumption that the trusted individual will not take advantage of one's vulnerability (McEvily, Perrone, & Zaheer, 2003, pp. 92-93).

Responsibility and accountability

Ryan and Stahl (2020, p. 74) point out that "End users should be able to justly trust AI organizations to fulfill their promises and to ensure that their systems function as intended […]. Building trust should be encouraged by ensuring accountability, transparency and safety of AI". "Criminal liability, the tort of negligence, and breach of warranty must be discussed before utilizing AI in medicine" (Matsuzaki, 2018, p. 268). Neri et al. (2020) pose the question of who is responsible for benefits and harms resulting from the use of AI in radiology, and, like Akinci D'Antonoli (2020), claim that radiologists remain responsible for the diagnosis when using AI, even if they might be validating something unknown that is based on black boxes and possible automation bias.

Sand et al. (2021) argues that the kind of accountability and responsibility that is being pursued in medical AI is connected to liability and blame. As an alternative, they propose a "forwardlooking responsibility," which "can be understood as a safeguard to decrease the risk of harm in cases of cognitive misalignment between the physicians and the AI system—when an AI output cannot be confirmed (verified or falsified)" (Sand et al., 2021, p. 3). Accordingly, the authors list the following responsibilities of clinicians: the duty to report uncertainty (sensitivity/specificity rates) to the patients; to understand and critically assess whether AI outputs are reasonable given a certain diagnostic procedure; to know and understand the input data and its quality; to have an awareness of their own experience and decline in skills; to have an awareness and understanding of the specificity of the task; and to assess, monitor, and report the output development over time.

Justice and fairness

Justice is one of the four principles of bioethics: autonomy, beneficence, non-maleficence, and justice (Beauchamp & Childress, 2001). Justice is also one of the three principles proposed in the Belmont Report (United States National Commission for the Protection of

Human Subjects of Biomedical Behavioral Research, 1978), one of the most widely recognized standards for biomedical ethics. Along with trust, transparency, accountability, and other principles, "diversity, non-discrimination and fairness" are principles that were proposed by the AI HLEG in 2018.

The association between injustice, discrimination, and unfair decisions made by AI systems has been also linked to bias in the reviewed literature, as "discrimination and unfair outcomes stemming from algorithms has become a hot topic within the media and academic circles" (Ryan & Stahl, 2020). Biased AI systems lead to unfair, discriminatory behavior or mistaken decisions (Morley et al., 2020) and to "algorithmic unfairness" (Abràmoff et al., 2020). Integrating AI systems in medicine incurs the risk of replicating discriminations that already exist in society; therefore, "the development of AI should promote justice while eliminating unfair discriminations, ensuring shareable benefits, and preventing the infliction of new harm that can arise from implicit bias" (Akinci D'Antonoli, 2020, pp. 508–509).

5 Empirical findings on trustworthy AI

As we have shown in our review of the scientific discourse, trust and trustworthiness has become a defining condition in the development and application of ethical AI. Whereas social scientific perspectives describe trust as fundamental for interpersonal interactions and society at large, approaches on trustworthy AI are addressing trust in AI mainly in terms of technical conditions and solutions. The social character of trust has received rather little attention and thus the interpersonal relationships, institutional conditions, and societal contexts that are of particular importance in health systems and settings. Trust plays a central role in situations of risk, vulnerability, and uncertainty – situations that are often encountered in the medical context. Sociological approaches claim that trusting relationships and trust in an institution, e.g., the healthcare system, is determined by the trust in its representatives, e.g., doctors, and, vice versa, trust in the medical system provides the legitimacy of its experts in providing a trustworthy environment (Meyer, Ward, Coveney, & Rogers, 2008).

In following the concept of trust that is constituted by several trustor/trustee relationships, such as trust in persons (e.g., scientists who trust each other, patients who trust scientists and clinicians), technology, and institutions (Bijker et al., 2016; Wyatt et al., 2013), we identified several understandings of trust in/trustworthiness of AI in our empirical study, which we will highlight in the next sections.

As we have observed in our literature review, we also found in the empirical material that trust or trustworthiness is often put in relation to explainability and interpretability. The emphasis is less focused on the technical inner-workings of an AI system, but more on the knowledge about the criteria by which AI reaches a decision. What kind of knowledge is seen as needed to build trust in AI is connected to the question of who the users of these systems are and what level of detail is necessary for them. As one of the interview partners puts it, radiologists and "people in the middle" (e.g., radiographers) need to have a workable but critical understanding of the system that is relevant to their practices.

> "[…] maybe you should be aware of, you know, how the decisions are made at least. So you will be able to explain it to a patient or at least to detect when something is

> wrong, right? […] I think it's quite interesting to sort of like provide, you know, […] explaining what it does. […] And also it's very important to say that the […] end user shouldn't trust 100% the results. […] And actually the AI results should be accompanied with some explanation. And here we talk about explainability. You know, to make sure the decision is sort of like a backup with something that makes sense. Because it has been in many cases that software got the right answer, but because of the wrong sort of like a purpose." Engineer

The understanding of how the system works does not necessarily require additional qualification by medical professionals but can be arranged in interdisciplinary institutional arrangements, as the following interview partner highlights:

> "[…] for implementing these tools, for trusting them, if we don't understand something to have someone that may go to the code and let the physician know why it's working like this, or if it doesn't work, what's going on, we need also engineers in the radiology departments." Academic radiologist

As we have seen in the reviewed scientific literature, trust is repeatedly conflated with other terms, such as acceptance and explainability (and the latter is often mixed with or replaced by, transparency, accountability, reliability, etc.), often without giving clear definitions or clarifying the boundaries between the concepts. In the empirical material, we found that professionals also use different terms such as accuracy or credibility to characterize their expectations towards AI systems:

> "I'm curious to know what is the accuracy of the system." Radiologist

> "[…] trust in AI directly relates to credibility of AI tools. […] it means that I can depend on the tool and be sure that it will do exactly what was promised by the AI developers." Clinician

This conceptual variety reflects the expectations towards explainability in broader terms but defines a common intention. Instead of opening the "black box" in order to see the technical decision-making process, the expectation is the traceability of the decision process as proof that AI is trustworthy as a tool, that it fulfils its purpose and that the outcome is intelligible to the medical professional, as it can be illustrated with the following quotes:

> "The problem of this approach using data-driven models, and algorithms in research, is that usually our systems that cannot be explained by themselves. There is no explanation about the results. Probably there, if the process is well done, probably you would achieve a good performance of the algorithm in place, but you don't know the reasons why the algorithms propose this one thing or another. This is one important point, no? And medical science is based on trust. And if they cannot explain what's the reason of the recommendations, the trust can be put to doubt." Physician

> "If I were a doctor, I would rather have a model that gives a confidence interval. Just instead of only giving a behind the scenes. Also catching all the errors and mistakes, all the adversarial attacks. If you can detect it, of course it builds more trust." Engineer

Trust in AI is trust "in the making". To be convinced of the systems and its accuracy also means that trust must grow over time based on experiences.

> "The doctors still trust the imaging, right? But there is no way of validating that these images are actually from what is inside the human body. Why do they still trust?

> Why do they trust scanners? […] For scanners, maybe in the beginning people were not trusting when the MRI scans came. […] Now we cannot live without it. AI is going to become like that, right? In the beginning nobody trusted. All these mechanisms are added on top of it, explainability, interpretability, uncertainty estimations. Maybe we will get there when they start trusting AI the same way they trust now the MRI scanners. It takes time." Data scientist

As mentioned earlier, trust is considered as situated at the interplay between individual and system. Following the empirical data, the trustworthiness of an AI application could be informed by the professional community, e.g., that AI performs as good as the "average radiologist", the scientific system, e.g., where the accuracy of AI is demonstrated in high impact factor papers, or the expert, in particular the doctor, and their experience and judgement, e.g., in comparing their own interpretation of an image with the output of an AI system. Furthermore, trust or trustworthiness of AI is also situated in the interplay of scientific practice (in terms of rational reasoning based on facts, proofs and evidence) and uncertainty, which is a fundamental characteristic of medical science. "Gut feelings" as discursive proxy for the key role emotions play in all this are therefore crucial for conceptualizing trustworthy AI that is predominantly defined by technical aspects.

> "I just know about the feelings of not feeling confident about using something. […] Sometimes many things in medicine have happened almost by gut feelings." Academic radiologist

Gut feelings or intuition are informed by past experiences and, as has also been shown in our previous research (Goisauf & Durnová, 2018), are also essential for research participants and patients in making decisions. In this context, the human factor is particularly evident as trust is needed to manage uncertainty – and this also refers to the accountability of human actors in dealing with the uncertainty that AI might create and the responsibility that comes with it:

> "The problem is that the input we are going to enter is very subjective. The final diagnosis is also subject to subjectivity, so we don't have 100% guarantee that the final diagnosis is correct. […] So everything is very subjective. So we are turning a system with subjectivity. […] essentially nobody will take responsibility of a software. If a software is wrong, who is responsible for this failure? I prefer to assume my risks, assuming that I'm a human and my opinion may be wrong, then delegating my responsibility to a software. Essentially, this is the main problem of AI." Radiologist

The agency of the medical profession towards the machine mentioned therein is also reflected in the understanding of patients and the role of AI.

> "I mean, imaging has changed the paradigm of diagnosis. Starting from mammography and MRIs and- Who can deny that? So if we can add something to help the physicians, so if the machines can learn from many cases, and that is the database purpose, where the lesion lays, it's interesting. But should we leave the machine performing by itself? I don't think so." Patient representative

A possible interplay between patient, doctor and machine is clearly outlined by the patient representative:

> "I would like to see a very strong AI capable of spotting illness where the doctors cannot see. Very powerful tools but guided by doctors. With the clinician eye between the results and the patient." Patient representative

From a sociological perspective, trust is particularly functional in managing uncertainty and knowledge gaps. Hence, the relationship between trust and uncertainty in AI needs closer examination. Following this conceptualization, the simplified causality that more knowledge leads to more trust is a misconception as complete knowledge would render trust obsolete. What we can conclude from the empirical data so far is that the knowledge of uncertainty and the question of which knowledge is meaningful for which users to foster agency and accountability is of high importance.

6 Guidelines on trustworthy AI

As shown, the existing body of literature on the 'Ethics of AI' in healthcare contains a variety of principles put forth from various viewpoints and with distinct methodologies, which may lead to the development of either broad or specific guidelines. For instance, the FUTURE-AI initiative has selected guiding principles drawn from the accumulated experiences and best practices from five large European projects on AI in medicine. These guiding principles to achieve trustworthy AI in the specific context of medical imaging are (Lekadir et al., 2021):

- Fairness,
- Universality,
- Traceability,
- Usability,
- Robustness, and
- Explainability.

Principle-based approaches can be enriched by perspectives that take into consideration the context in which the proposed principles operate (Mittelstadt, 2019). Instead of proposing a new set of principles and guidelines, we build on the FUTURE-AI principles and elaborate on relevant ethical and social/societal aspects in the following comments. Our research suggests that it would be beneficial to focus on:

- The broader cultural and societal context in which AI is being designed and implemented, as this can shape ethical considerations and impact the acceptability and deployability of AI technologies.
- The potential power dynamics between healthcare providers and patients, as AI may exacerbate existing inequalities or create new ones.

By taking these contextual factors into account, we can develop nuanced and comprehensive ethical frameworks for the use of AI in healthcare. Based on our research, we thus recommend directing attention towards the following aspects.

Consideration of potential biases, in particular sex and gender dimensions

Research suggests that using imbalanced datasets that do not include enough data from minority groups to train deep-learning-based systems in healthcare can affect the performance of pathology classification for those groups (Larrazabal, Nieto, Peterson, Milone, & Ferrante, 2020). Additionally, social categories like gender and socioeconomic status can

influence diagnosis and potentially result in missed detection of diseases like breast cancer (Rauscher, Khan, Berbaum, & Conant, 2013). It is important to consider how identity traits can impact the application of AI systems in healthcare to avoid producing skewed datasets that could harm certain minority people and groups.

The results of our research indicate that it is important to pay close attention to potential biases in the training and validation data of AI in order to produce fair algorithms, in particular to sex and gender dimensions in datasets, rather than ignoring them (Larrazabal et al., 2020; Tannenbaum, Ellis, Eyssel, Zou, & Schiebinger, 2019). Some researchers (Cirillo et al., 2020; Pot, Kieusseyan, & Prainsack, 2021) suggest introducing a desirable bias to counteract the effects of undesirable biases that can lead to unintended or unnecessary discrimination. Furthermore, based on our empirical research, it could be considered that synthetic data can be used as a data augmentation strategy to balance datasets and evaluate AI application in different contexts, so that undesirable biases can be reduced or even avoided.

There are already efforts to introduce the sex and gender perspective in biomedical research[10] (Oertelt-Prigione, Parol, Krohn, Preissner, & Regitz-Zagrosek, 2010; Oertelt-Prigione & Regitz-Zagrosek, 2012). As the project "Gendered Innovations" suggests[11], it is equally important to question the background assumptions that circulate unquestioned within biomedical research community. One of these assumptions is that sex and gender are binary categories (male and female; men and women, respectively). Key institutions drive calls for a more complex and inclusive approach to sex and gender. For instance, the American Medical Association updated its policies in 2018 to affirm the medical spectrum of sex and gender, by stating that "sex and gender are more complex than previously assumed. […] It is essential to acknowledge that an individual's gender identity may not align with the sex assigned to them at birth. A narrow limit on the definition of sex would have public health consequences for the transgender population and individuals born with differences in sexual differentiation, also known as intersex traits" (American Medical Association, 2018). Also, in the latest version of the International Statistical Classification of Diseases and Related Health Problems (ICD-11), gender identity-related health was redefined from gender "disorder" to "incongruence" and reclassified as "sexual health" (World Health Organization, 2019). One of the leading medical journals, The Lancet, updated their author guidelines[12], encouraging the enrolment of women and ethnic groups into clinical trials and to analyze the data accordingly, considering influences and associations of sex and gender. When it comes to data collection, there are efforts to provide recommendations[13] which highlight the need for data collectors to receive training on the categories of sex and gender, and to apply that knowledge in the creation of questions designed to capture the diversity of sexes and genders (Colaço & Watson-Grant, 2021).

[10] See: https://inb-elixir.es/events/train-trainer-integration-sex-and-gender-dimension-life-sciences-research, accessed 08.09.2023.

[11] See: http://genderedinnovations.stanford.edu/methods/concepts.html, accessed 07.09.2023

[12] See https://www.thelancet.com/pb/assets/raw/Lancet/authors/tlqh-info-for-authors-1664975851987.pdf, accessed 05.09.2023.

[13] See: https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-sets/mental-health-services-data-set/submit-data/data-quality-of-protected-characteristics-and-other-vulnerable-groups/gender-identity, accessed on 08.09.2023.

In the field of AI in medicine, most algorithms are designed without considering the impact of sex and gender on health and disease differences in individuals (Cirillo et al., 2020). Nonetheless, there are efforts to incorporate the sex and gender dimensions in AI applications in medicine and biomedicine (Cirillo, Solarz, & Guney, 2022).

The FUTURE AI guidelines provide advice on collecting data, whenever possible, on different relevant attributes (see Fairness 2), rather than ignoring them. It is important to take these attributes into account when it comes to identifying potential biases. The purpose of the collection of this type of data is to minimize bias, and it should be accompanied by an especial attention to non-discrimination measures and by an effort to minimize the risk of re-identification.

We have focused here on sex and gender as two of the attributes that are being more widely addressed in a critical way. In Europe, other categories such as race and ethnicity are not available for data collection due to historical reasons. When this data is missing, it is not only missing data but it is something that speaks about certain aspects of our society that also need to be considered. When it comes to data collection, building balanced databases, and assessing AI tools for potential biases, this lack of data poses a challenge.

Interdisciplinarity and embedded ethics approach

Embedded ethics refers to the integration of ethical considerations into the design and development process of technology. It involves identifying potential ethical issues and addressing them proactively to minimize harm and promote ethical behavior. In this sense, it is important to introduce interdisciplinary knowledge of societal issues from the beginning of the development of AI applications (McLennan et al., 2020) and throughout all stages of design, development and implementation. Researchers and providers often do not have the expertise to identify or address structural factors. For instance, learning to identify biases can promote "algorithmic fairness," and ML approaches might be used to correct them (Abràmoff et al., 2020). An interdisciplinary approach that takes into consideration different aspects involved in the development and application of AI, as well as different approaches to understanding the intertwinement between AI and societal issues, may contribute to build trust in AI applications.

Researchers working in the field of ethics of AI in medicine will need to strive for accuracy and precision by providing clear definitions for concepts such as trustworthiness or trust in a specific context and by situating them within a broader context of societal issues. In order to do this, interdisciplinary research with social scientists but also with clinicians to incorporate clinical concepts (Lekadir et al., 2021) will be crucial. This connects with the FUTURE AI guidelines (see General 5), which emphasize the need to investigate ethical, social, and societal issues in interdisciplinary teams that engage in discussions regarding these issues throughout the project. Engagement and ongoing information flows are necessary to be generated in order for these conversations to be fruitful for the AI applications.

Situated practices

It is often assumed that data is objective and neutral, but data is always embedded in a particular context. This means that data is always influenced by the assumptions and biases of the person or group collecting and analyzing the data. Therefore, it is crucial to question these background assumptions and consider the context in which the data was collected to arrive at a more accurate and nuanced understanding of the data. This relates to Fairness 1 from the FUTURE AI recommendations, which states that it is crucial to define the source of bias, as bias in medical AI is application specific. It is important to pay attention to the specific context of application and the specific group attributes of the population where the AI tool is applied.

Attention to context, including a deep understanding of pre-existing inequalities and vulnerable groups, is also important in order to implement AI in real-world situations. Recommendation Universality 1 refers to the need to define the specific clinical settings here the technology will be applied at the design phase, while University 4 claims that AI tools should be evaluated for their local clinical validity and should perform well in the local population. Universality 4 also points out a crucial factor: AI should fit well in the existing local clinical flow. As shown by the empirical data, doctor-patient relationships are crucial when it comes to trust; there are trust relationships that have already been established and even institutionalized and AI applications need to build on these relationships and fit into these institutional dynamics.

To do that, it is important to assess different specific factors that may affect the AI tool's performance in real-world situations (recommendations Robustness 1, 2, and 3). If an AI tool is trained on data that is representative of the real-world variations encountered in clinical practice, it is more probable that clinicians, citizens, and other stakeholders will have trust in its use.

Recommendation Usability 3 also connects the usability of the tool with the engagement with relevant stakeholders when it claims that AI tools should be evaluated for usability in the real-world with representative and diverse end-users with different attributes (regarding sex, gender, age, clinical role, digital proficiency or (dis)ability). Furthermore, this recommendation is in line with the principle of diversity that has been proposed by the AI – HLEG in their guidelines from trustworthy AI.

Stakeholder engagement

As shown by our empirical research, keeping humans in the loop or proposing an approach that privileges human oversight (recommendation traceability 6) is paramount to building trust in AI applications. Furthermore, to ensure that the development and implementation of AI technologies are effective and beneficial, it is crucial to define the relevant stakeholders and involve them from the early stages of the research process (linked to recommendation Usability 1). This includes interdisciplinary teams, patients, and other healthcare professionals. Meaningful involvement is key, which means creating a space for stakeholders to interact with each other and engage in knowledge-building exercises. It is important to define what needs to be known by different stakeholders (linked to recommendation Explainability 1).

It is often stated that for trust to be gained, there should be extensive knowledge about the inner workings of algorithms and AI applications, and that the existence of more data will lead to more trust. Nonetheless, it is a matter of engaging relevant stakeholders in knowledge-building exercises in which information can be shared and needs can be identified. What is relevant for different stakeholders is context- and user-dependent.

By involving all relevant stakeholders in decision-making, we can ensure that AI technologies are developed and implemented in ways that are both effective and ethical.

A look into the future: environmental aspects

The energy consumption of training large neural networks and running high-performance computing systems can be significant, leading to increased greenhouse gas emissions. Additionally, the production and disposal of electronic components used in AI systems can contribute to electronic waste and other environmental (Wu et al., 2022).

It is important for AI researchers and developers to consider the environmental impacts of their work and strive to minimize any negative effects. This can be done through more energy-efficient algorithms and hardware, as well as responsible sourcing and disposal of electronic components.

Although the environmental well-being is one of the principles for trustworthy AI proposed by the AI – HLEG, the literature on the topic is scarce. Our observation is that this is an important topic that needs further research and more attention in the future.

The following table is intended as a toolbox that translates the presented guidelines into practical questions.

Table 1: Toolbox for the guidelines on trustworthy AI

Guideline	Takeaways	Considerations
Inclusion of sex and gender	Gender bias and sex assumptions in biomedical research can have an adverse impact in the lives of patients. For this reason, it is important to include the categories of sex and gender biomedical research. This inclusion can help raise awareness on: - How new concepts and theories of gender can bring to light new evidence. - How background assumptions about sex and	- Do we, as researchers, assume certain gender roles? - Are we aware of possible gender biases in our field? - Do we assume that sex and gender are binary? - Do we take into consideration the interactions between sex and gender? (Adapted from Gendered Innovations)

	gender can shape concepts and theories in the field.	- Do we understand the difference between sex and gender and whether it is relevant to work with these categories in the particular domain the AI tool will be applied?
Interdisciplinarity	Interdisciplinary perspectives may be necessary to fully understand and address the impact of research at different, intertwined levels (scientific, societal, economic, legal, etc.).	- Have we considered collaborating with researchers from different disciplines? - Are there any potential interdisciplinary connections between our research topic and other fields of study? - Have we identified any potential knowledge gaps that could be addressed by integrating multiple disciplines? - Have we taken steps to ensure that all team members feel valued and included, regardless of their disciplinary background - Have we created a space where people from different backgrounds (I.e. data scientists and radiologists) can talk to each other about the tools that are being created? - Have we understood the local institutional dynamics of the place where the AI tool will be applied?
Embedded ethics	Integration of ethical considerations into the design and development process of medical AI. Identification of potential ethical issues and proactive approach to minimize harm and promote ethical behavior.	- Have we involved diverse stakeholders in the design and planning of our research project? - Have we identified and addressed potential ethical issues that may arise in our research project? - Have we considered the potential long-term implications of our research findings in the communities and individuals involved?

		- Have we established a mechanism for ongoing ethical review and monitoring of our research project?
Situated practices		- Have we questioned our background assumptions in order to identify sources for potential bias? - Have we considered the specific group attributes of the population where the AI tool will be applied? - Have we understood pre-existing inequalities and potentially vulnerable groups in the context where the AI tool will be applied? - Have we evaluated for the local clinical validity of our AI tools, and assess whether they will perform well in the local population?
Stakeholder Engagement	Early definition of relevant stakeholders, including interdisciplinary teams and patients. Involvement of all relevant stakeholders in decision-making.	- Who are the stakeholders involved in the project, and what are their specific needs, knowledge, and interests? - How will the AI tool impact each stakeholder group, and what are their potential concerns? - What is the level of understanding of the AI technology among each stakeholder group? - How can each stakeholder group be involved in knowledge-building exercises, and how can their feedback be incorporated into the project? - Have we made efforts to disseminate our research findings to relevant stakeholders, including research participants

		and the wider community?
Environmental issues		- Have we considered using renewable energy sources to power the AI models and algorithms? - How can the energy consumption of our AI models be minimized? - Have we considered the environmental impact of the data centers where our AI models and algorithms are hosted? - Have we developed a plan for responsibly disposing of electronic waste generated by our research group/proposal?

7 Summary of further resources

Systematic review: Ethics of AI in Radiology: A Review of Ethical and Societal Implications

Goisauf, M., & Cano Abadía, M. (2022). Ethics of AI in Radiology: A Review of Ethical and Societal Implications. *Frontiers in Big Data, 5*(850383). doi:10.3389/fdata.2022.850383

Webinar: Ethics of AI in Imaging: Ethical and Societal Implications

Artificial Intelligence (AI) applications in medicine are hoped to improve healthcare and to advance health equity. While AI carries the potential to improve health services, its ethical and societal implications need to be carefully considered to avoid harmful consequences for individuals and groups, especially for the most vulnerable. It is therefore inevitable to identify what types of ethical issues are raised by AI, and to analyze how these issues are tackled in biomedical research. In this webinar, we give an overview of the results of a comprehensive and systematic review of academic literature as well as workshop outcomes. They problematise approaches such as 'trustworthy AI' and 'explainable AI' that shape the ethics discourse on AI. The webinar concludes with a reflection on the topics identified that shape the understanding of 'Ethics of AI' and the gaps in the discourse.

Project: Gendered Innovations

The peer-reviewed Gendered Innovations project:

1) develops practical methods of sex, gender, and intersectional analysis for scientists and engineers;
2) provides case studies as concrete illustrations of how sex, gender and intersectional analysis leads to innovation.

Paper: Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare

Cirillo, D., Catuara-Solarz, S., Morey, C., Guney, E., Subirats, L., Mellino, S., . . . Mavridis, N. (2020). Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *npj Digital Medicine, 3*(1), 81. doi:http://doi.org/10.1038/s41746-020-0288-5

Paper: FUTURE-AI: Guiding Principles and Consensus Recommendations for Trustworthy Artificial Intelligence in Medical Imaging

Lekadir, K., Osuala, R., Gallin, C., Lazrak, N., Kushibar, K., Tsakou, G., . . . Tsiknakis, M. (2021). FUTURE-AI: Guiding Principles and Consensus Recommendations for Trustworthy Artificial Intelligence in Medical Imaging. *arXiv preprint arXiv:2109.09658*. Retrieved from https://arxiv.org/abs/2109.09658

8 References

Abràmoff, M. D., Tobey, D., & Char, D. S. (2020). Lessons Learned About Autonomous AI: Finding a Safe, Efficacious, and Ethical Path Through the Development Process. *American Journal of Ophthalmology, 214*, 134-142. doi:https://doi.org/10.1016/j.ajo.2020.02.022

Akinci D'Antonoli, T. (2020). Ethical considerations for artificial intelligence: an overview of the current radiology landscape. *Diagnostic and interventional radiology, 26*(5), 504-511. doi:10.5152/dir.2020.19279

American Medical Association. (2018). AMA Adopts New Policies at 2018 Interim Meeting [Press release]. Retrieved from https://www.ama-assn.org/press-center/press-releases/ama-adopts-new-policies-2018-interim-meeting

Baier, A. (1986). Trust and antitrust. *Ethics, 96*(2), 231-260.

Balthazar, P., Harri, P., Prater, A., & Safdar, N. M. (2018). Protecting Your Patients' Interests in the Era of Big Data, Artificial Intelligence, and Predictive Analytics. *Journal of the American College of Radiology, 15*(3, Part B), 580-586. doi:https://doi.org/10.1016/j.jacr.2017.11.035

Beauchamp, T. L., & Childress, J. F. (2001). *Principles of biomedical ethics*. Oxford, New York: Oxford University Press.

Bijker, E. M., Sauerwein, R. W., & Bijker, W. E. (2016). Controlled human malaria infection trials: How tandems of trust and control construct scientific knowledge. *Social Studies of Science, 46*(1), 56-86. doi:10.1177/0306312715619784

Bjerring, J. C., & Busch, J. (2021). Artificial Intelligence and Patient-Centered Decision-Making. *Philosophy & Technology, 34*(2), 349-371. doi:10.1007/s13347-019-00391-6

Brady, A. P., & Neri, E. (2020). Artificial Intelligence in Radiology—Ethical Considerations. *Diagnostics, 10*(4), 231. Retrieved from https://www.mdpi.com/2075-4418/10/4/231

Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care—addressing ethical challenges. *The New England journal of medicine, 378*(11), 981-983. doi:10.1056/NEJMp1714229

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative research*. London/California/New Delhi: Sage.

Cirillo, D., Catuara-Solarz, S., Morey, C., Guney, E., Subirats, L., Mellino, S., . . . Mavridis, N. (2020). Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *npj Digital Medicine, 3*(1), 81. doi:10.1038/s41746-020-0288-5

Cirillo, D., Solarz, S. C., & Guney, E. (2022). *Sex and Gender Bias in Technology and Artificial Intelligence: Biomedicine and Healthcare Applications*. London: Academic Press, Elsevier.

Colaço, R., & Watson-Grant, S. (2021). *A Global Call to Action for Gender-Inclusive Data Collection and Use*. Retrieved from Research Triangle Park (NC):

Currie, G., Hawk, K. E., & Rohren, E. M. (2020). Ethical principles for the application of artificial intelligence (AI) in nuclear medicine. *European Journal of Nuclear Medicine and Molecular Imaging, 47*(4), 748-752. doi:10.1007/s00259-020-04678-1

Dubber, M. D., Pasquale, F., & Das, S. (2020). *The Oxford Handbook of Ethics of AI*. New York: Oxford University Press.

Ferretti, A., Schneider, M., & Blasimme, A. (2018). Machine learning in medicine: opening the new data protection black box. *Eur. Data Prot. L. Rev., 4*, 320.

Geis, J. R., Brady, A. P., Wu, C. C., Spencer, J., Ranschaert, E., Jaremko, J. L., . . . Kohli, M. (2019). Ethics of Artificial Intelligence in Radiology: Summary of the Joint European and North American Multisociety Statement. *Canadian Association of Radiologists Journal, 70*(4), 329-334. doi:10.1016/j.carj.2019.08.010

Goisauf, M., & Cano Abadía, M. (2022). Ethics of AI in Radiology: A Review of Ethical and Societal Implications. *Frontiers in Big Data, 5*(850383). doi:10.3389/fdata.2022.850383

Goisauf, M., & Durnová, A. P. (2018). From engaging publics to engaging knowledges: Enacting "appropriateness" in the Austrian biobank infrastructure. *Public Understanding of Science, 28*(3), 275-289. doi:10.1177/0963662518806451

High-Level Expert Group on AI. (2019). *Policy and Investment Recommendations for Trustworthy AI*. Retrieved from https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_policy_and_investment_recommendations.pdf

High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI*. Retrieved from Brussels: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Kaufman, J. (2013). Ethical dilemmas in statistical practice: the problem of race in biomedicine. In E. G. Laura & L. Nancy (Eds.), *Mapping 'race': critical approaches to health disparities research* (pp. 53–66). Ithaca: Rutgers University Press.

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine, 17*(1), 195. doi:10.1186/s12916-019-1426-2

Larasati, R., & DeLiddo, A. (2020). Building a trustworthy explainable AI in healthcare. In F. Loizides, M. Winckler, U. Chatterjee, J. Abdelnour-Nocera, & A. Parmaxi (Eds.), *Human Computer Interaction and Emerging Technologies: Adjunct Proceedings from the INTERACT 2019 Workshops* (pp. 209–214). Cardiff: Cardiff University Press.

Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H., & Ferrante, E. (2020). Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences, 117*(23), 12592-12594. doi:doi:10.1073/pnas.1919012117

Leavy, S. (2018). Gender bias in artificial intelligence: the need for diversity and gender theory in machine learning. *GE '18: Proceedings of the 1st International Workshop on Gender Equality in Software Engineering*, 14-16. doi:https://doi.org/10.1145/3195570.3195580

Lekadir, K., Osuala, R., Gallin, C., Lazrak, N., Kushibar, K., Tsakou, G., . . . Tsiknakis, M. (2021). FUTURE-AI: Guiding Principles and Consensus Recommendations for Trustworthy Artificial Intelligence in Medical Imaging. *arXiv preprint arXiv:2109.09658*. Retrieved from https://arxiv.org/abs/2109.09658

Lekadir, K., Quaglio, G., Tselioudis Garmendia, A., & Gallin, C. (2022). *Artificial Intelligence in Healthcare-Applications, Risks, and Ethical and Societal Impacts*. Retrieved from Brussels:

Li, J., Bzdok, D., Chen, J., Tam, A., Ooi, L. Q. R., Holmes, A. J., . . . Genon, S. (2022). Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity. *Science Advances, 8*(11). doi:10.1126/sciadv.abj1812

Matsuzaki, T. (2018). Ethical issues of artificial intelligence in medicine. *California Western Law Review, 55*(1), 255-273.

McEvily, B., Perrone, V., & Zaheer, A. (2003). Trust as an Organizing Principle. *Organization Science, 14*(1), 91-103. doi:10.1287/orsc.14.1.91.12814

McLennan, S., Fiske, A., Celi, L. A., Müller, R., Harder, J., Ritt, K., . . . Buyx, A. (2020). An embedded ethics approach for AI development. *Nature Machine Intelligence, 2*(9), 488-490. doi:https://doi.org/10.1038/s42256-020-0214-1

Meyer, S., Ward, P., Coveney, J., & Rogers, W. (2008). Trust in the health system: An analysis and extension of the social theories of Giddens and Luhmann. *Health Sociology Review, 17*(2), 177-186. doi:10.5172/hesr.451.17.2.177

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence, 1*(11), 501-507. doi:10.1038/s42256-019-0114-4
Morley, J., Machado, C. C. V., Burr, C., Cowls, J., Joshi, I., Taddeo, M., & Floridi, L. (2020). The ethics of AI in health care: A mapping review. *Social Science & Medicine, 260*, 113172. doi:10.1016/j.socscimed.2020.113172
Murtagh, M. J., Minion, J. T., Turner, A., Wilson, R. C., Blell, M., Ochieng, C., . . . Burton, P. R. (2017). The ECOUTER methodology for stakeholder engagement in translational research. *BMC medical ethics, 18*(1), 24. doi:10.1186/s12910-017-0167-z
Neri, E., Coppola, F., Miele, V., Bibbolino, C., & Grassi, R. (2020). Artificial intelligence: Who is responsible for the diagnosis? *La radiologia medica, 125*(6), 517-521. doi:10.1007/s11547-020-01135-9
Noble, S. U. (2018). *Algorithms of oppression. How search engines reinforce racism*. New York: New York University Press.
O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown.
Oertelt-Prigione, S., Parol, R., Krohn, S., Preissner, R., & Regitz-Zagrosek, V. (2010). Analysis of sex and gender-specific research reveals a common increase in publications and marked differences between disciplines. *BioMed Central Medicine Medical Research Methodology, 8*, 70-80. doi:http://doi.org/10.1186/1741-7015-8-70
Oertelt-Prigione, S., & Regitz-Zagrosek, V. (Eds.). (2012). *Sex and Gender Aspects in Clinical Medicine*. London: Springer.
Owens, K., & Walker, A. (2020). Those designing healthcare algorithms must become actively anti-racist. *Nature Medicine, 26*(9), 1327-1328. doi:10.1038/s41591-020-1020-3
Pesapane, F., Volonté, C., Codari, M., & Sardanelli, F. (2018). Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States. *Insights into Imaging, 9*(5), 745-753. doi:10.1007/s13244-018-0645-y
Pot, M., Kieusseyan, N., & Prainsack, B. (2021). Not all biases are bad: equitable and inequitable biases in machine learning and radiology. *Insights Imaging, 12, 13*. doi:https://doi.org/10.1186/s13244-020-00955-7
Quinn, T. P., Jacobs, S., Senadeera, M., Le, V., & Coghlan, S. (2021). The three ghosts of medical AI: Can the black-box present deliver? *Artificial Intelligence in Medicine*, 102158. doi:https://doi.org/10.1016/j.artmed.2021.102158
Rasheed, K., Qayyum, A., Ghaly, M., Al-Fuqaha, A., Razi, A., & Qadir, J. (2022). Explainable, Trustworthy, and Ethical Machine Learning for Healthcare: A Survey. *Computers in Biology and Medicine, 149*(1-26). doi:https://doi.org/10.1016/j.compbiomed.2022.106043
Rauscher, G. H., Khan, J. A., Berbaum, M. L., & Conant, E. F. (2013). Potentially missed detection with screening mammography: does the quality of radiologist's interpretation vary by patient socioeconomic advantage/disadvantage? *Annals of Epidemiology, 23*(4), 210–214. doi:https://doi.org/10.1016/j.annepidem.2013.01.006
Roberts, D. E. (2008). Is Race-Based Medicine Good for Us?: African American Approaches to Race, Biomedicine, and Equality. *Journal of Law, Medicine & Ethics, 36*(3), 537-545. doi:10.1111/j.1748-720X.2008.302.x
Ryan, M., & Stahl, B. C. (2020). Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. *Journal of Information, Communication and Ethics in Society, 19*(1), 61-86. doi:10.1108/JICES-12-2019-0138
Sand, M., Durán, J. M., & Jongsma, K. R. (2021). Responsibility beyond design: physicians' requirements for ethical medical AI. *Bioethics, 36*(2), 1–8. doi:https://doi.org/10.1111/bioe.12887

Schiebinger, L., & Schraudner, M. (2011). Interdisciplinary approaches to achieving gendered innovations in science, medicine, and engineering 1. *Interdisciplinary Science Reviews, 36*(2), 154–167.

Tannenbaum, C., Ellis, R. P., Eyssel, F., Zou, J., & Schiebinger, L. (2019). Sex and gender analysis improves science and engineering. *Nature, 575*(7781), 137-146. doi:10.1038/s41586-019-1657-6

Terry, G., Hayfield, N., Clarke, V., & Braun, V. (2017). Thematic analysis. In C. Willig & W. Stainton Rogers (Eds.), *The SAGE handbook of qualitative research in psychology* (Vol. 2, pp. 17-37). London: Sage.

Tizhoosh, H. R., & Pantanowitz, L. (2018). Artificial intelligence and digital pathology: challenges and opportunities. *Journal of pathology informatics, 9*(38), 1-6. doi:10.4103/jpi.jpi_53_18

United States National Commission for the Protection of Human Subjects of Biomedical Behavioral Research. (1978). *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research*. Retrieved from New York:

Ware, A. B. (2018). Algorithms and Automation: Fostering Trustworthiness in Artificial Intelligence. *Honors Theses and Capstones, 416*, 1-37.

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., & Jung, K. (2019). Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine, 25*(9), 1337-1340. doi:https://doi.org/10.1038/s41591-019-0548-6

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., . . . Bourne, P. E. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data, 3*. doi:10.1038/sdata.2016.18

World Health Organization. (2019). Gender incongruence and transgender health in the ICD. Retrieved from https://www.who.int/standards/classifications/frequently-asked-questions/gender-incongruence-and-transgender-health-in-the-icd

World Health Organization. (2021). *Ethics and governance of artificial intelligence for health: WHO guidance*. Geneva: World Health Organization

Wu, C.-J., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., . . . Hazelwood, K. (2022). *Sustainable AI: Environmental implications, challenges and opportunities.* Paper presented at the 5th MLSys Conference, SantaClara, CA, USA.

Wyatt, S., Harris, A., Adams, S., & Kelly, S. E. (2013). Illness Online: Self-reported Data and Questions of Trust in Medical and Social Research. *Theory, culture & society, 30*(4), 131-150. doi:10.1177/0263276413485900