Accelerating Implementation of Artificial Intelligence in Radiotherapy through Explainability*

Luca Heising[1,2,*,†]

[1]Department of Economics and Management (TiSEM), Tilburg University, Tilburg, NL
[2]Department of Radiation Oncology (Maastro), GROW School for Oncology and Reproduction, Maastricht University Medical Centre+, Maastricht, NL

Abstract

To enhance the radiotherapy workflow, many artificial intelligence (AI) applications have been proposed. To date, only a limited number of the proposed AI applications have been implemented into clinical practice. Lack of trust is often mentioned as the limiting factor due to the inherent black-box characteristics of AI. Explainable AI (xAI) methods are being introduced as tool to alleviate the lack of trust in these non-transparent systems. To study the effect that xAI has on clinicians' trust, a survey was developed and distributed. Preliminary findings conclude that clinicians do not necessarily mistrust AI, yet, they seem to find transparency important. xAI could serve as a shared mental model (SMM) between the clinician and AI to maximize human-AI collaboration. Future work will look at the role that xAI plays in SMMs and how xAI must be designed to fully exploit AI for radiotherapy whilst remaining safe and ethical.

Keywords

Radiotherapy, Implementation, Healthcare, Trust, xAI design, Shared Mental Models

1. Introduction

In previous years, artificial intelligence (AI) has gained its place in the spotlight to optimize healthcare processes [1]. The promise of AI is large, with many applications proposed to aid clinicians in their work, among which in radiotherapy. The recent advancements in AI can be attributed, in part, to Deep Learning (DL) algorithms, a subset of AI that utilizes multiple hidden layers to discern patterns from input data. Deep Learning allows for accurate pattern recognition in large datasets, engendering the opportunity to optimize healthcare workflows and provide aid in labor-intensive tasks. In the radiotherapy workflow, AI has been proposed for every step of the workflow. Examples of radiotherapy-related AI applications relatively high on the technology readiness level scale are automated delineation of anatomic contours and tumors, automated treatment planning systems, and medical imaging quality improvement algorithms.

Late-breaking work, Demos and Doctoral Consortium, colocated with The 1st World Conference on eXplainable Artificial Intelligence: July 26–28, 2023, Lisbon, Portugal

*Corresponding author.
†These authors contributed equally.
l.m.heising@tilburguniversity.edu (L. Heising)
https://research.tilburguniversity.edu/en/persons/luca-heising (L. Heising)
0000-0002-9426-2858 (L. Heising)

One of the applications where AI has been proven to be a promising solution is treatment error detection to facilitate adaptive radiotherapy (ART) [2], which is especially interesting for a group of lung cancer patients (limited-stage small cell lung cancer) as they respond well to radiotherapy treatment. Yet, the implementation of AI is lagging behind, partly because AI has not been fully embraced by clinicians as AI poses additional risks [3]. For example, due to the hidden layers of deep neural networks, the outcome of the AI can be difficult to understand and/or interpret, and thus challenging to verify. This phenomenon is often described as the black box; what goes into the network and what comes out can be observed, but not what happens inside and why.

In addition to hurdles in risks and regulation, developers and clinicians are currently often working in silos, which can result in a delay in the implementation of relevant AI tools in clinical practice as they do not meet the exact needs of the clinician or the clinician is unaware of the full potential of these tools. Despite the challenges posed by AI, being overly careful with its development and implementation can lead to underuse and missing out on cost- and quality-improvement opportunities. To bridge the gap between development and implementation and to enable the safe implementation of transparent AI in radiotherapy, the current PhD project investigates how xAI can contribute to safe and ethical implementation of AI whilst fully exploiting its opportunities for radiotherapy. The end product of the project aims to deliver guidelines on how to maximize collaborative performance through design of xAI in radiotherapy.

2. Research motivation

Radiotherapy is a highly quantitative field in healthcare which requires high precision. The use of technology is high, and professionals in this field are used to computational modelling. For this reason, radiotherapy presents an appropriate setting to implement AI as it is expected that the skepticism towards AI is relatively low compared to other fields in healthcare. The introduction of AI could potentially optimize workflows, increase efficiency, and allow for standardization of care, which is requested as the work pressure is increasing and world health organization expects a deficit of over 10 million healthcare workers by 2030. As of today, AI applications have been proposed for every step of the radiotherapy workflow, yet, the implementation is falling behind. Amongst other barriers, the black box characteristics [4, 5] and lack of trust in AI [6, 7, 8] are mentioned as important barriers to the implementation of AI. This is especially troublesome for the small group of small cell lung cancer patients that depend on AI as the adaption is mostly significant to this group of patients. xAI aims at opening up this black box and at the same time increasing trust in AI, fostering the AI-clinician collaboration. Design of xAI asks for a different approach than the design of regular AI algorithms as it serves as a bridge between the algorithm and the user. The literature is scarce on how the design of this translation between AI and humans can best be done and how this may differ between different user groups.

3. Key related work that frames your research

3.1. Key xAI concepts and definitions

Although the literature wields different definitions for concepts related to explainable AI (xAI), we distinguish three main concepts; explainability, interpretability, and transparency, where explainability and interpretability both contribute to the AI's transparency.

The AI model is explainable if it is clear how the model came to its prediction. A decision tree can, thus, be considered more explainable than for example a deep neural network. The outcome is interpretable if it is clear why the model outputs a certain prediction, for example, by providing information on which features were most important for the model's prediction. A model is considered transparent if it is by itself understandable [9]. Yet, these definitions do not reveal anything about the quality of the xAI method, which is based on the user and relates to human-ai interaction. Different layers to consider in xAI are philosophical foundations of what an explanation is, social attribution of the explanation, the cognitive processes underpinning how people explain and evaluate explanations, and social explanations or how people communicate their explanations [10].

3.2. Trust in AI

Trust plays a large role in the adoption of technology as it affects both perceived usefulness and perceived ease of use [11], the two main drivers of technology acceptance (TAM). Many methods proposing xAI or implementing xAI argue that lack of trust is one of the main barriers for AI adoption and that opening up the black box that is AI should ameliorate the lack of trust [12, 13, 14]. Trust is, however, a complex concept which is influenced by many factors. It is yet unclear how xAI plays a role in trust or lack thereof.

Trust is a fundamental aspect in the decision making process [15]. Yet, trust is an ill-defined concept with different meanings in different disciplines. The meaning of trust in computing is defined as follows: "Estimated subjective probability that an entity exhibits reliable behavior for particular operation(s) under a situation with potential risks" . Several factors play a role in the outcome of trust such as belief, experience, rationality, uncertainty, reliability, and more [15]. Recently, a conceptual model of trust, perceived risk and reliance on AI was proposed [16]. This framework lays out different actors of trust in an AI decision aid. Transparency of AI is, however, not covered by the proposed framework. Previous work found that increasing AI transparency concurrently increases trust in AI [17], [18]. Our work argues that transparency mediates the feeling of control over AI. The amended conceptual model on trust in AI, including AI transparency (Figure 1). In the regarding PhD research, we study the effect of AI transparency on perceived control and on trust in AI from clinicians in radiotherapy and radiology.

3.2.1. Perceived trustworthiness

Perceived trustworthiness is determined by many different characteristics, including specifications of the AI itself. This includes the input, process and outcome of the AI, where the outcome can be represented in the form of an AI metric such as accuracy. Other factors affecting the

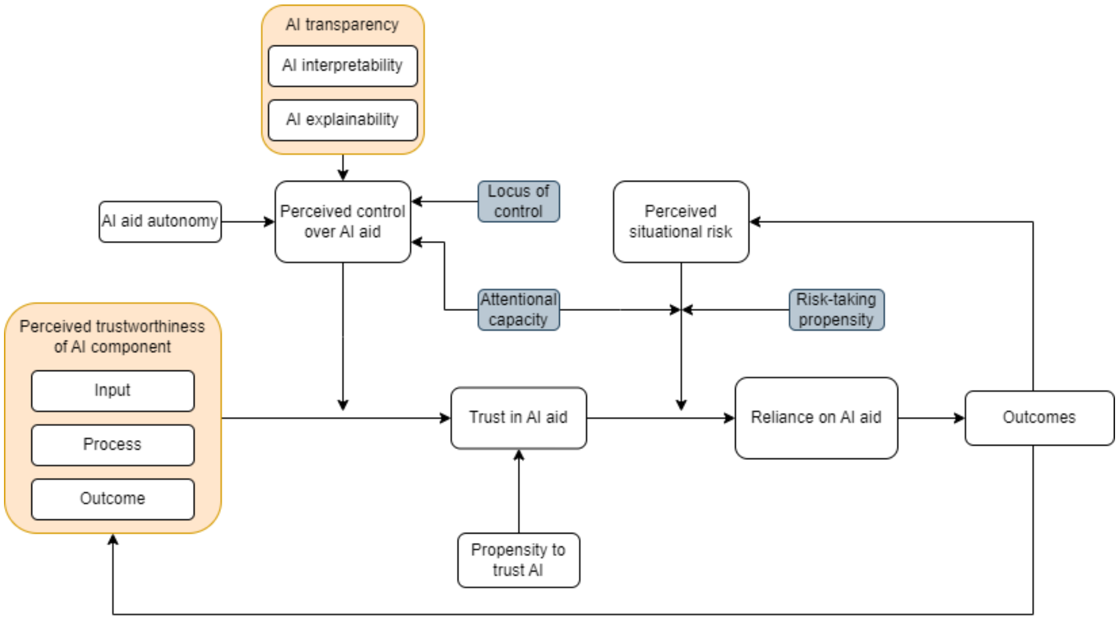

Figure 1: Conceptual model of trust in artificial intelligence and the role of xAI

perceived trustworthiness relate to external factors such as organizational context, personal identity, and many more.

3.2.2. Perceived control

According to the previously proposed conceptual framework on trust, the trust in AI is moderated by the perceived control over AI [16]. xAI allows the user to maintain human oversight influences users' perceived control over AI. This was, however, not yet tested before.

In literature, in the majority of studies, trust is discussed in the context of under-trust. On the other hand, overly trusting AI can potentially lead to harmful outcomes as well. Inappropriate trust in AI could potentially lead clinicians in the wrong direction and a major drawback of AI its confidence in its predictions. In the long run, a brain drain may occur as clinicians are not required to perform the tasks themselves anymore. The lack of knowledge might result in situations where the user is not capable of evaluating the AI output.

With this framework, we examine trust in AI from a perspective of under trust, it is not clear what the impact of too much trust in AI can have. Future work also aims to look at the impact of too much trust in AI and how xAI can play a role finding the equilibrium between over- and under-trust.

3.3. xAI design

Mental models are knowledge structures to describe, explain and predict the world around them [19]. Shared mental models (SMM) refer to how individual tasks and team tasks are described, explained and predicted by members of that team. The quality of a shared mental model is related to team performance. For human-AI collaboration, an SMM be helpful in increasing the team performance of human-AI collaboration. According to Andrews et al. (2023) [19], xAI can serve as a method for SMM formation and elicitation.

To establish human-centered design, close collaboration with users from the beginning of the design process is required [20]. For traditional technology, the technology readiness level reveals information about the maturity of certain technology and what steps must be considered for implementation into practice. More recently, a similar scale was proposed to establish the technology readiness level of machine learning algorithms [21] This is a ten step workflow starting at first principles, which is purely research based, up to the last level: deployment.

4. Specific research questions, hypothesis and objectives

With the PhD research we aim to investigate how xAI can contribute to the safe implementation of artificial intelligence for a synergistic collaboration between professionals in radiotherapy and AI. As multiple user-groups in radiotherapy will use AI, we will research how AI can contribute to all of these groups and how the communication in the form of xAI may differ between the different users. To reach these goals, the following research questions were drawn up for the start-up phase of the PhD:

How is trust in AI affected by xAI from the perspective of radiologist and radiotherapist?

What are opportunities and challenges regarding xAI for radiotherapy? (systematic literature review)

How can xAI serve as a mental model for human-centric AI design for different user groups in radiotherapy from a design perspective?

5. Research approach, methods, and rationale for testing the research hypothesis

This PhD research aims at studying how xAI can contribute to the implementation and adoption of AI in radiotherapy.

5.1. Survey on trust in AI

To study trust in AI, extension of the conceptual framework for trust in AI is made. We argue that xAI allows for higher perceived control over the AI, thereby increasing trust in AI. To measure this, we developed a survey and distributed the survey amongst all professionals in radiotherapy and radiology through newsletters and snowballing. After two months, the number of responses saturated at 206 respondents.

For every construct of the conceptual framework, we developed several survey questions. All questions were validated by performing a pilot study with professionals from every participant group. The content validity was measured using the content validity index (CVI), the participants from the pilot study where asked to fill out a 5-point likert scale on how relevant that specific question was. In total, ten people participated in the pilot study, of which five radiation oncologists, two RTTs, two clinical therapists and one radiologist. Questions receiving a 2 or lower on the CVI were removed or adjusted. The data will be analyzed using structural equation modelling.

5.2. Systematic literature review on xAI in radiotherapy

The second research question will be answered by performing a systematic literature review. Inclusion criteria for the search strategy included all deep learning application in the radiotherapy workflow that applied some sort of xAI method. In total, the search led to 420 results, and after filtering according to the PRISMA flow diagram ended up with 180 articles.

5.3. Development explainable AI framework

For the development of an explainable AI framework, we stick to the design science research approach [22]. Stakeholders of the dose-guided radiotherapy workflow are Radiotherapy oncologists (RTO), Medical physicists (MP), Lab technicians (RTT), and AI Developers. Each of these groups has a different set of tasks and responsibilities requiring different information. A large part of the DGRT workflow is laborious due to the large false positive rates. The information used to assess the dose differences include the PDIs and CBCT. Although less interpretable, the PDI holds more information as this is a direct measurement of the dose given to the patient and allows for the detection of machine-related treatment errors besides positioning errors and anatomical errors visible in the CBCT.

A focus group was set up to commonly establish the automation potential of the DGRT workflow and the degree of augmentation per user group. Three separate sessions were held: one with MPs, one with RTOs, and one with RTTs. During the session, the participants pointed out where the workflow bottlenecks occur and prioritized these bottlenecks. Further steps include a five-day design sprint with professionals in the field to design the xAI solution and establish the requirements. The sprint will be followed by a feedback round with the individual user groups.

6. Results and contributions to date

To date, the survey on trust in AI has been developed and distributed. As mentioned before, 206 respondents filled out the survey. The first analyses revealed that trust in AI does not seem to be lacking in radiotherapy, on the contrary, most respondents seem to have high trust in AI. Yet, interpretability of AI is not unwelcome, which would suggest that the xAI does not only serve as a trust generator. These findings should be interpreted cautiously as the analyses are not completed or based on testing.

7. Expected next steps and final contribution to knowledge

The next steps for the PhD study include developing an xAI governance framework for data governance, xAI development and deployment, and (x)AI implementation and adoption. Additionally, the guidelines which will be derived from the systematic literature review will be tested in a clinical setting for error detection in dose-guided radiotherapy.

References

[1] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, Y. Wang, Artificial intelligence in healthcare: past, present and future, Stroke and vascular neurology 2 (2017).

[2] C. J. Wolfs, R. A. Canters, F. Verhaegen, Identification of treatment error types for lung cancer patients using convolutional neural networks and epid dosimetry, Radiotherapy and Oncology 153 (2020) 243–249. URL: https://www.sciencedirect.com/science/article/pii/S0167814020308239. doi:https://doi.org/10.1016/j.radonc.2020.09.048, physics Special Issue: ESTRO Physics Research Workshops on Science in Development.

[3] O. Asan, A. E. Bayrak, A. Choudhury, Artificial intelligence and human trust in healthcare: focus on clinicians, Journal of medical Internet research 22 (2020) e15154.

[4] C. Fiorino, R. Jeraj, C. H. Clark, C. Garibaldi, D. Georg, L. Muren, W. van Elmpt, T. Bortfeld, N. Jornet, Grand challenges for medical physics in radiation oncology, Radiotherapy and Oncology 153 (2020) 7–14.

[5] L. Arbelaez Ossa, G. Starke, G. Lorenzini, J. E. Vogt, D. M. Shaw, B. S. Elger, Re-focusing explainability in medicine, Digital health 8 (2022) 20552076221074488.

[6] F. Cabitza, A. Campagner, C. Balsano, Bridging the "last mile" gap between ai implementation and operation:"data awareness" that matters, Annals of translational medicine 8 (2020).

[7] L. Petersson, I. Larsson, J. M. Nygren, P. Nilsen, M. Neher, J. E. Reed, D. Tyskbo, P. Svedberg, Challenges to implementing artificial intelligence in healthcare: a qualitative interview study with healthcare leaders in sweden, BMC Health Services Research 22 (2022) 1–16.

[8] F. Gama, D. Tyskbo, J. Nygren, J. Barlow, J. Reed, P. Svedberg, Implementation frameworks for artificial intelligence translation into health care practice: scoping review, Journal of Medical Internet Research 24 (2022) e32215.

[9] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information fusion 58 (2020) 82–115.

[10] A. Angelopoulou, E. Kapetanios, D. H. Smith, V. Steuber, B. Woll, F. Zeller, Explanation in human-ai systems, Frontiers in Artificial Intelligence 5 (2022) 1048568.

[11] P. A. Pavlou, Consumer acceptance of electronic commerce: Integrating trust and risk with the technology acceptance model, International journal of electronic commerce 7 (2003) 101–134.

[12] L. Strohm, C. Hehakaya, E. R. Ranschaert, W. P. Boon, E. H. Moors, Implementation of artificial intelligence (ai) applications in radiology: hindering and facilitating factors, European radiology 30 (2020) 5525–5532.

[13] S. Reddy, S. Allan, S. Coghlan, P. Cooper, A governance model for the application of ai in health care, Journal of the American Medical Informatics Association 27 (2020) 491–497.

[14] C. L. Brouwer, A. M. Dinkla, L. Vandewinckele, W. Crijns, M. Claessens, D. Verellen, W. van Elmpt, Machine learning applications in radiation oncology: Current use and needs to support clinical implementation, Physics and imaging in radiation oncology 16 (2020) 144–148.

[15] J.-H. Cho, K. Chan, S. Adali, A survey on trust modeling, ACM Computing Surveys (CSUR) 48 (2015) 1–40.

[16] E. Solberg, M. Kaarstad, M. H. R. Eitrheim, R. Bisio, K. Reegård, M. Bloch, A conceptual model of trust, perceived risk, and reliance on ai decision aids, Group & Organization Management 47 (2022) 187–222.

[17] E. Glikson, A. W. Woolley, Human trust in artificial intelligence: Review of empirical research, Academy of Management Annals 14 (2020) 627–660.

[18] D. Shin, The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai, International Journal of Human-Computer Studies 146 (2021) 102551.

[19] R. W. Andrews, J. M. Lilly, D. Srivastava, K. M. Feigh, The role of shared mental models in human-ai teams: a theoretical review, Theoretical Issues in Ergonomics Science 24 (2023) 129–175.

[20] H. Chen, C. Gomez, C.-M. Huang, M. Unberath, Explainable medical imaging ai needs human-centered design: guidelines and evidence from a systematic review, NPJ digital medicine 5 (2022) 156.

[21] A. Lavin, C. M. Gilligan-Lee, A. Visnjic, S. Ganju, D. Newman, S. Ganguly, D. Lange, A. G. Baydin, A. Sharma, A. Gibson, et al., Technology readiness levels for machine learning systems, Nature Communications 13 (2022) 6039.

[22] A. Hevner, S. Chatterjee, A. Hevner, S. Chatterjee, Design science research in information systems, Design research in information systems: theory and practice (2010) 9–22.