J. Mantas et al. (Eds.)

doi:10.3233/SHTI230409

The Representation of Trust in Artificial Intelligence Healthcare Research

Jan-Oliver KUTZA[a,b,1], Niels HANNEMANN[a], Ursula HÜBNER[b], Birgit BABITSCH[a]

[a] *Department of New Public Health, Osnabrück University, Nelson-Mandela-Straße 13, D 49076 Osnabrück, Germany*

[b] *Health Informatics Research Group, Osnabrück University of Applied Sciences, Albrechtstraße 30, D 49076 Osnabrück, Germany*

Abstract. Artificial intelligence (AI) tends to emerge as a relevant component of medical care, previously reserved for medical experts. A key factor for the utilization of AI is the user's trust in the AI itself, respectively the AI´s decision process, but AI-models are lacking information about this process, the so-called Black Box, potentially affecting user´s trust in AI. This analysis´ objective is the description of trust-related research regarding AI-models and the relevance of trust in comparison to other AI-related research topics in healthcare. For this purpose, a bibliometric analysis relying on 12985 article abstracts was conducted to derive a co-occurrence network which can be used to show former and current scientific endeavors in the field of healthcare based AI research and to provide insight into underrepresented research fields. Our results indicate that perceptual factors such as "trust" are still underrepresented in the scientific literature compared to other research fields.

Keywords. Artificial Intelligence, Trust, Black-Box, bibliometric analysis

1. Introduction

Data driven artificial intelligence (AI) systems are based on models in a specific application domain to achieve a specific purpose [1], for example to support medical decision-making [2]. Extending their capabilities, they are reaching out into areas previously reserved for medical experts [3] with comparable accuracy [4]. An important factor for the utilization of AI-models, is the user´s trust in such models [5], defined "*as the reliance by an agent that actions prejudicial to their well-being will not be undertaken by influential others*" [6]. Thus, when the decision process enabled by AI-models is transparently illustrated, users tend to be more likely to trust AI-models [7] encouraging them to utilize such models [8]. However, approaches in machine learning such as convolutional neural networks tend to lack extensive information about how they generate decisions, resulting in a users' limited ability to understand and trust the decisions [9-10;1]. This phenomenon has been called the Black Box problem [11;1]. Due to the relevance of trust for the acceptance and utilization of AI enabled systems, it is necessary to better understand what trust means in the context of AI and how it is

[1] Corresponding Author: Jan-Oliver Kutza, Department of New Public Health, Osnabrück University, Nelson-Mandela-Straße 13, D 49076 Osnabrück, Germany. Email: jkutza@uni-osnabrueck.de.

developed. This study therefore aims to (i) identify studies that address AI in healthcare, (ii) determine the extent to which trust is a research interest in the context of these research efforts, and (iii) present the relative scientific relevance of trust compared to other AI-related topics.

2. Methods

For the study´s purpose, a bibliometric analysis was conducted [12] using Gephi version 0.10 [13]. Publications were searched in Scopus, PubMed, Cochrane Library, Web of Science, CINAHL, APA PsycInfo, Philosopher´s Index, and Sociological Abstracts. The search terms groups two major topics. Topic no. 1 refers to methods collectively known as AI, including: "Artificial intelligence", "xai", "Machine Learning", "ml", "xml", "neural network", "nn", "Deep Learning", "dl", "Block Chain". Topic no. 2 covers the domains of health and healthcare, including: "health", "health care", "healthcare". Within the topics, relevant terms were linked with the Boolean operator *OR*, and the topics were linked via the operator *AND*. Articles were included when at least one term from each topic was mentioned in the title and abstract and at least one term from topic no. 1 was mentioned in the article´s keywords or MeSH terms. Articles, written in English and published between the years 2013 and 2023 were included. Data collection took place on March 10, 2023. A total of 15,885 articles were identified with 12,985 remaining after title and abstract based deletion of duplicates. To generate a network file for Gephi, the dataset was then imported into the nocodefunctions [14] application with the parameters being adjusted to only search for English terms, a minimum length of words of two and all other parameters on default settings. The derived network file incorporated a total of 1,999 nodes and over 1.5 million edges and was imported into Gephi. Manual data aggregation was then performed independently by two researchers, removing irrelevant nodes like "alcohol" or "pipeline" and edges and aggregating semantically consistent nodes, like "artificial intelligence" and "AI" or "wearable", "wearable device" and "wearable sensor". Inconsistencies were resolved by joint decisions. In this way, edges were reduced to 2,771 and nodes to 79, including 54 consolidated nodes. The number of terms of the aggregated nodes were summed up. For illustration, the *force atlas* layout was chosen and run until sufficient stability was achieved. The *repulsion strength* was set to 1800.0, the *attraction strength* was set to 50.0, and the *automatic stabilization* function and *sizing* were enabled. In addition, the *noverlap* and *label adjust* functions were executed. Due to node consolidation, several self-loop edges were introduced and suppressed using the *self-loop* filter. By calculating and applying the *average weighted degree*, the color of the nodes corresponds to the average weighted number of incoming edges per node, with lighter colors corresponding to a higher number of connected mentioned terms. The total number of terms mentioned corresponds to the size of each node. Edges were colored by assigning a higher weight to the lighter colors. The network shown uses an *edge thickness* of 5.0, an *opacity* of 90.0 and *curved edges* were deactivated, with all other options based on the default settings. For visualization purposes, slight manual adjustments were made to the placement of the top and bottom nodes towards the center of the network.

3. Results

The resulting network (Fig. 1) shows that the presented terms differ strongly in number. From 22,327 ("health") to 122 ("participation") with frequent mentions of the terms "artificial intelligence" (22,140) (Tab. 1), "machine learning" (17,584) and "neural network" (11,888) and rarely mentioned terms such as "fairness" (133), "black box" (136) and "trust" (543). The weight of the undirected edges ranges from 8.616 ("artificial intelligence" to "health") to 0.0024 ("usability" to "participation"). "Health" (8.616), "effectiveness" (1.8378), and "classification" (1.6651) have the highest weight associated with the "Artificial Intelligence" node, while "fairness" (0.002355) or "explainable AI" (0.0079) show a comparatively low weight.

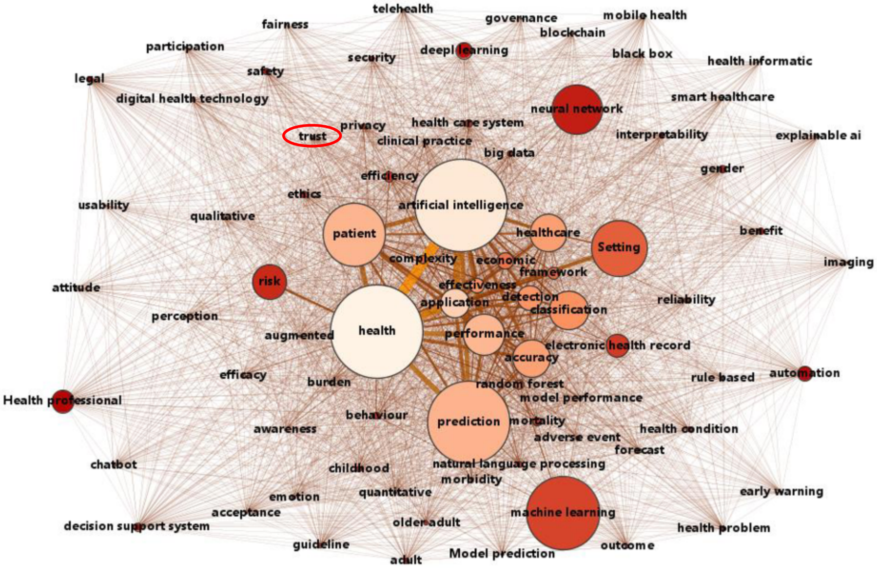

Figure 1. Co-occurrence network AI in Healthcare (trust circled in red, upper left quadrant).

Table 1. Representation of trust research in AI-models in healthcare

start node	amount	target node	amount	Edge weight
artificial intelligence	22140	trust	543	0.2795
trust	543	machine learning	17584	0.0047
neural network	11888	trust	543	0.0055
explainable ai	643	trust	543	0.0306
trust	543	black box	136	0.0212

4. Discussion

With the help of the present co-occurrence network, it is possible to depict which research areas have already been addressed to a greater or lesser extent in the field of artificial intelligence in healthcare. It can be seen that process- and outcome-oriented research areas, which can be assigned to the concepts of "performance," "accuracy," "recognition," or "cost-effectiveness," for example, have been researched to a much greater extent than those that are more on the psychosocial, ethical and legal level. These include, for example, concepts such as "emotion", "trust", "fairness", ethics" or "legal". There are limitations associated with this analysis. One limitation relates to the fact that only English-language articles were included in the analysis, leading to language bias and an underestimation of relevant research in other languages. Incorporating abstracts only in the dataset, the co-occurrence network addressed trust as a central research focus. Including full text articles could lead to broader result space. Since an inductive merging and deletion of terms took place in the course of data preparation for better readability and interpretability, slight deviations from the basic data set may appear in the number of individual head terms mapped in the network.

5. Conclusion

Although the scientific literature points to the importance of trust in AI enabled systems for their use, we could show that the relevance of trust-related research is underrepresented in current AI-related research in healthcare compared to other, more outcome-oriented research topics. It seems useful to pursue systematic approaches to gain a more fundamental understanding of the trust-building process when using AI based applications. These approaches should include, for example, research on user characteristics, attitudes, or on features provided by the AI application that may influence the development of trust. Reflecting potential multi-layered properties trust and its building processes may carry, combining qualitative and quantitative methods like Delphi studies and experimental study designs in form of RCTs would be advisable.

References

[1] Poole DI, Goebel RG, Mackworth AK. Computational intelligence. A Logical Approach. New York: Oxford University Press; 1998. 1-22 pp, ISBN: 978-0-19-510270-3.

[2] Adadi A, Berrada M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence. IEEE Access. 2018; 6: 52138-52160. doi: 10.1109/ACCESS.2018.2870052.

[3] Yu KH, Beam L, Isaac S. Artificial intelligence in healthcare. Nature biomedical engineering. 2018;2(10):719-731, doi: https://doi.org/10.1038/s41551-018-0305-z.

[4] Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future healthcare journal. 2019;6(2):94-98. doi: 10.7861/futurehosp.6-2-94.

[5] Kirste M. Augmented Intelligence – Wie Menschen mit KI zusammenarbeiten. iit-Themenband Künstliche Intelligenz Technologie | Anwendung | Gesellschaft In: Wittpahl V, editor. Springer Vieweg; 2019. p. 58-71. ISBN: 978-3-662-58042-4

[6] Hancock PA, Billingd DR, Schaefer KE Can you trust your robot?. Ergonomics in Design. 2011;19(3):24-29. doi: 10.1177/1064804611415045.

[7] Wang W, Siau K. Artificial intelligence, machine learning, automation robotics, future of work and future of humanity: A review and research agenda. Journal Database Management (JDM). 2019;30(1):61-79. doi: 10.4018/JDM.2019010104.

[8] Du M, Liu N, Hu X. Techniques for interpretable machine learning. Communications of the ACM. 2019;63(1):68-77. doi: 10.1145/3359786.

[9] Fernandez A, Herrera F, Cordon O, del Jesus MJ, Marcelloni F. Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to? IEE Computational intelligence magazine. 2019;14(1):69-81. doi: 10.1109/MCI.2018.2881645.

[10] Gunning D, Stefik M, Choi M, Miller T, Stumpf S, Yang GZ. XAI – Explainable artificial intelligence. Science robotics. 2019;4(37):eaay7120. doi: 10.1126/scirobotics.aay7120.

[11] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence. 2019;1(5):206-215. doi: https://doi.org/10.1038/s42256-019-0048-x.

[12] Khudzari JMD, Kurian J, Tartakovsky B, Raghavan GSV. Bibliometric analysis of global research trends on microbial fuel cells using Scopus data. Biochemical engineering journal. 2018;136:51-60. doi: https://doi.org/10.1016/j.bej.2018.05.002.

[13] Bastian M, Heymann S, Jacomy M. Gephie: an open source software for exploring and manipulating networks. Proceedings of the Third International AAAI Conference on Weblogs and Social Media; 2009 May 17-20; San Jose, CA: PKP Publishing Services Network. p. 361-362. doi: https://doi.org/10.1609/icwsm.v3i1.13937.

[14] Levallois, C. Nocodefunctions; 2023; https://nocodefunctions.com/index.html; accessed March 17, 2023.