Ali Shiri
University of Alberta, Edmonton, Alberta, Canada

Imagining Future Human-centred AI

Abstract or Résumé:
The rapid growth in the development of AI-enabled and intelligent systems and services calls for more serious and urgent attention to human-focused and human-centred principles, methods, and approaches. This paper reports on a study of recent human-centred AI research literature to examine the extent to which the three principles of s*afety*, *trustworthiness*, and *reliability* have been addressed in recent AI publications.

1. Introduction

Recent AI research has been paying increasing attention to the human aspects of AI system developments. Shneiderman's new, thought-provoking, and widely reviewed book on Human-Centered AI (2022) clearly articulates a set of key principles that can guide the development of future responsible AI-enabled systems. He argues that "human-Centered AI shows how to make successful technologies that amplify, augment, empower, and enhance human performance". He further emphasizes the key principles of s*afety*, *trustworthiness*, and *reliability* as particularly important for the development of any AI-enabled systems that aim to be human-centred.

Furthermore, recent AI research has addressed such key principles as contextual morality for artificial intelligence that identifies opportunities for future work through a FACT-based (Fairness, Accountability, Context and Transparency) perspective (van Berkel et al., 2022) and the trustworthiness of robots and autonomous systems with respect to safety, security, health, human -machine interaction and ethics (He et al., 2022). The European Union has made a set of policy recommendations stressing that "in order to succeed, the EU must realign its AI-related initiatives and focus them on mission-based innovations—that is, large-scale projects to develop human-centred AI so that it augments our intelligence in a computer–human symbiosis to solve the societal problems of our time" rather than AI systems that supersede humans (Carriço, 2018). An important institutional development related to human-centred AI is the formation of the Stanford University Human-centered AI institute established in 2019 to help imagine how AI will impact humans in the future and to guide build the future of AI. The institute has three major areas of research, namely human impact, augmented human capabilities, intelligence.

With the rapid increase in the number of publications on AI in the past decade, from 16,999 records in 2012 to 38608 in November 2022, it is timely and important to examine the extent to which the AI reported research takes into account the human-centred aspects of AI. Given the importance of human-centred approaches to the development of AI-enhanced systems, this study aims to explore the extent to which the three key principles of *safety*, *trustworthiness*, and *reliability*, proposed by Shneiderman, have appeared, addressed, or discussed in the published AI research literature. We argue that the publications that have included a discussion of all of the three key principles mentioned above should provide a reasonable set of scholarly works that

represent some of the key themes, topics, and challenges that the designers of AI-enabled systems are currently facing. Based on this argument, this study will more specifically identify the frequently used themes and topics that have co-occurred with Shneiderman's three key principles of *safety*, *trustworthiness*, and *reliability.* This kind of analysis will afford us the opportunity to examine both the emerging nature of human-centred AI discussion and the trending themes, topics, and terms in the related literature.

2. Methods and tools

This study made use of metadata analytics methods using publication metadata records from the Scopus multidisciplinary database as well as text analytics and visualization tools.

Database

Considering the increasing interdisciplinarity of the domain of artificial intelligence and the growth of journal articles, conference proceedings, and research reports that address AI in a variety of disciplinary domains, it was decided to use the Scopus multidisciplinary database for this study. We used the time period 2012-2022 for this study as it represents a decade of rapid increase in the number of AI publications. The Scopus interdisciplinary database was used to conduct searches on human-centred AI and the variations of the three keywords that reflect the three human-centred AI principles proposed by Shneiderman, namely s*afety*, *trustworthiness*, and *reliability.* All searches for metadata records containing the above keywords were conducted on November 24, 2022. The Scopus database allows for searches to be carried out on a number of metadata elements, including titles, abstracts, author keywords, and index keywords. The search delimiters allow for a broadcast search across these textual metadata elements. This kind of search retrieves all the keywords in the following fields: title, abstract, author keywords, and index keywords. Detailed search strategies for these keywords are provided below. The metadata elements that were examined for analysis in this study included the following:

- title
- author keywords
- indexed keywords
- abstract
- publication date
- publication source

Analytical methods and tools

Given the small size of the publication metadata records, both automatic and manual analysis of author keywords and index keywords were conducted. For the analysis of abstracts only automatic text analysis was utilized. Two text analysis and visualization tools were used, namely WordStat 8, content analysis and text mining software (part of QDA miner) and Voyant, an open source web-based text analysis and visualization tool. Specific term analysis techniques that were used include term frequency, co-occurrence analysis, and trend analysis.

Search strategy

As noted above searches were conducted on AI and the variations of the three keywords that reflect the three human-centred AI principles proposed by Shneiderman, namely s*afety*, *trustworthiness*, and *reliability.* Our actual search strategy can be seen in the following:

Artificial intelligence OR AI (533, 697 records):
Artificial intelligence OR AI (limit search to 20122-2022)
Narrow down search by the following keywords:
Safe OR Safety
Reliable or Reliability
Trustworthy or trustworthiness

Search strategy	**Number of records**
(TITLE-ABS-KEY ("artificial intelligence") OR TITLE-ABS-KEY (ai))	533, 697
(TITLE-ABS-KEY (*"artificial intelligence"*) OR TITLE-ABS-KEY (*ai*)) AND (LIMIT-TO (PUBYEAR , *2022*) OR LIMIT-TO (PUBYEAR , *2021*) OR LIMIT-TO (PUBYEAR , *2020*) OR LIMIT-TO (PUBYEAR , *2019*) OR LIMIT-TO (PUBYEAR , *2018*) OR LIMIT-TO (PUBYEAR , *2017*) OR LIMIT-TO (PUBYEAR , *2016*) OR LIMIT-TO (PUBYEAR , *2015*) OR LIMIT-TO (PUBYEAR , *2014*) OR LIMIT-TO (PUBYEAR , *2013*) OR LIMIT-TO (PUBYEAR , *2012*))	344, 302
(TITLE-ABS-KEY ("artificial intelligence") OR TITLE-ABS-KEY (ai) AND TITLE-ABS-KEY ((safe OR safety)))	16, 582
(TITLE-ABS-KEY ("artificial intelligence") OR TITLE-ABS-KEY (ai) AND TITLE-ABS-KEY ((trust OR trustworthy OR trustworthiness)))	6, 193
(TITLE-ABS-KEY ("artificial intelligence") OR TITLE-ABS-KEY (ai) AND TITLE-ABS-KEY ((safe OR safety)) AND TITLE-ABS-KEY ((reliable OR reliability)))	1,916
(TITLE-ABS-KEY ("artificial intelligence") OR TITLE-ABS-KEY (ai) AND TITLE-ABS-KEY ((trust OR trustworthy OR trustworthiness)) AND TITLE-ABS-KEY ((reliable OR reliability)))	627
(TITLE-ABS-KEY ("artificial intelligence") OR TITLE-ABS-KEY (ai) AND TITLE-ABS-KEY ((safe OR safety)) AND TITLE-ABS-KEY ((trust OR trustworthy OR trustworthiness)))	526
(TITLE-ABS-KEY ("artificial intelligence") OR TITLE-ABS-KEY (ai) AND TITLE-ABS-KEY ((safe OR safety)) AND TITLE-ABS-KEY ((trust OR trustworthy OR trustworthiness)) AND TITLE-ABS-KEY ((reliable OR reliability)))	110

Table 1. Search strategy for human-centred AI, safety, trustworthiness, reliability

Rationale for the choice of keywords

It should be noted that the above terms were specifically and deliberately selected to be searched for this study as they represent the exact match keywords used by Shneiderman in his work along with the noun forms of the terms. The intention was not to create a list of synonymous or semantically related terms for the above keywords as Shneiderman has used these terms consistently and purposefully to make strong arguments about the future of more human-centred and responsible AI enhanced systems. Therefore, terms such as machine learning and its specific techniques (supervised learning, unsupervised learning, semi-supervised learning, deep learning, or transfer learning) have not been included. Similarly, the specific areas of AI such as computer vision, natural language processing, robotics, cognitive computing are not included in the search.

3. Preliminary findings and observations

Analysis of data

Given that the extracted data consisted of metadata records from the Scopus database, we report here the analysis of the key metadata elements, namely title, date, publication source, abstract, index terms, and author terms. Figure 1 shows the publication trends for human-centred AI research that has incorporated the discussion of safety, trustworthiness, reliability. As can be seen, an increase is observable starting from the year 2019, suggesting increased attention and treatment of these principles in AI research.

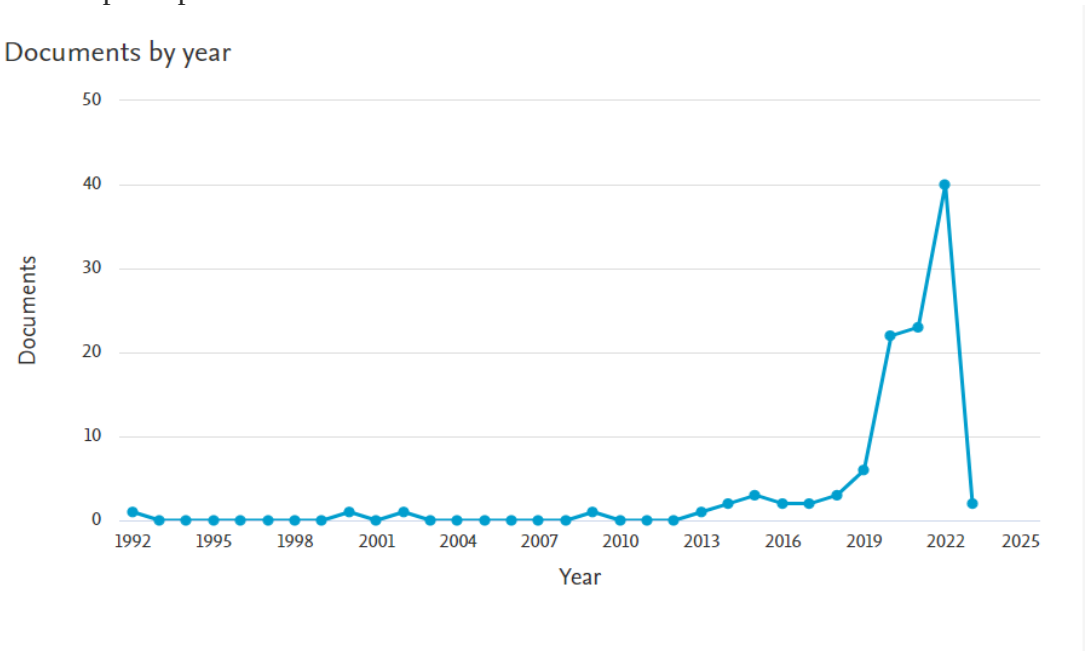

Figure 1. Publication trends on human-centred AI research addressing of safety, trustworthiness, and reliability

Index term analysis

A phrase frequency analysis of the Index terms shows some interesting topics and themes. Apart from the high frequency terms such as artificial intelligence, machine learning and deep learning, there are a number of terms that refer to Shneiderman's proposed principles. Safety engineering, accident prevention, safety critical, user experience, human computer interaction, behaviorual research, autonomous driving, data reliability, human control, risk assessment, human robot interaction, fatigue detection, privacy preservation, trust management and trust mechanism, and uncertainty analysis.

Author keyword analysis

For the analysis of author keywords both manual and automatic text analysis methods were used. Authors attributed particular importance to the notion of explainability. In AI research, explainability is considered one of the most frequently used terms in recent publications as it reflects the importance of being able to explain what data, algorithms, methods and approaches a researcher is using as they develop systems or try to identify trends or behavioural patterns. There are a set of important considerations that guide the development of explainable AI, including transparency, causality, bias, fairness, and safety (Hagras, 2018). In addition to trust, reliability and safety, the authors made use of such terms as accountability, explainability, misbehaviour and intrusion detection, interpretability, learning (deep learning, reinforced learning, experiential learning, continual learning, machine learning), usability and accessibility, ethical and responsible AI, bias, risks, security, Human control and human-AI interaction (assisted living, aging, computational pathology, telemedicine, cancer detection, medical device regulation, diagnostic), privacy preservation and public policy, autonomous lethal weapons, and autonomous cars.

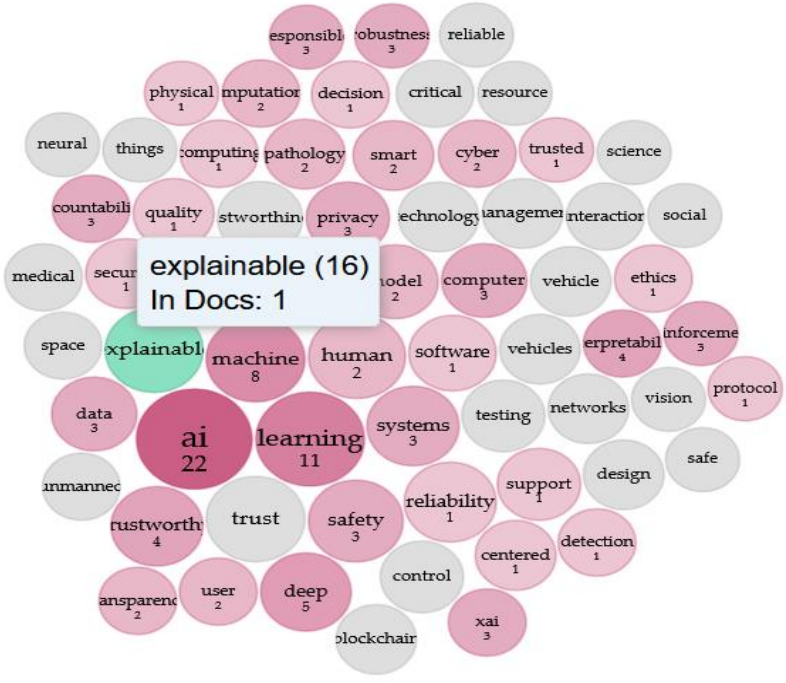

Figure 2. Author keyword co-occurrence analysis

Abstract keyword analysis

Voyant was used for text analysis. Among the most frequently used terms were trust, safety and safe, data, learning, human, reliable and reliability, trustworthy and trustworthiness, autonomous, control, security, explainability, communication, privacy, transparency, accuracy, healthcare, and self driving cars. Among the top 20 terms are: trust, trustworthy, safety, reliability, data, information, human, learning, followed by vehicles, use, security, safety, quality, explainability, and decision.

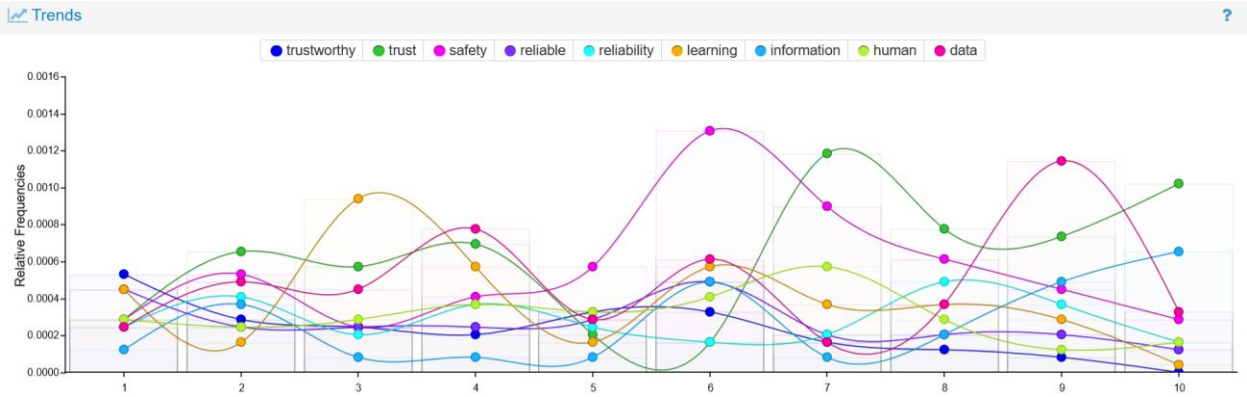

Figure 3. Analysis of high frequency terms in abstracts of AI publications

Publication source analysis

As noted earlier, there were 110 publications that were analyzed for this project. We conducted a manual categorization of the sources of publications to specifically see if there are observable trends in the disciplinary focus of sources. Figure 4 shows an overview of the publication sources based on their disciplinary focus.

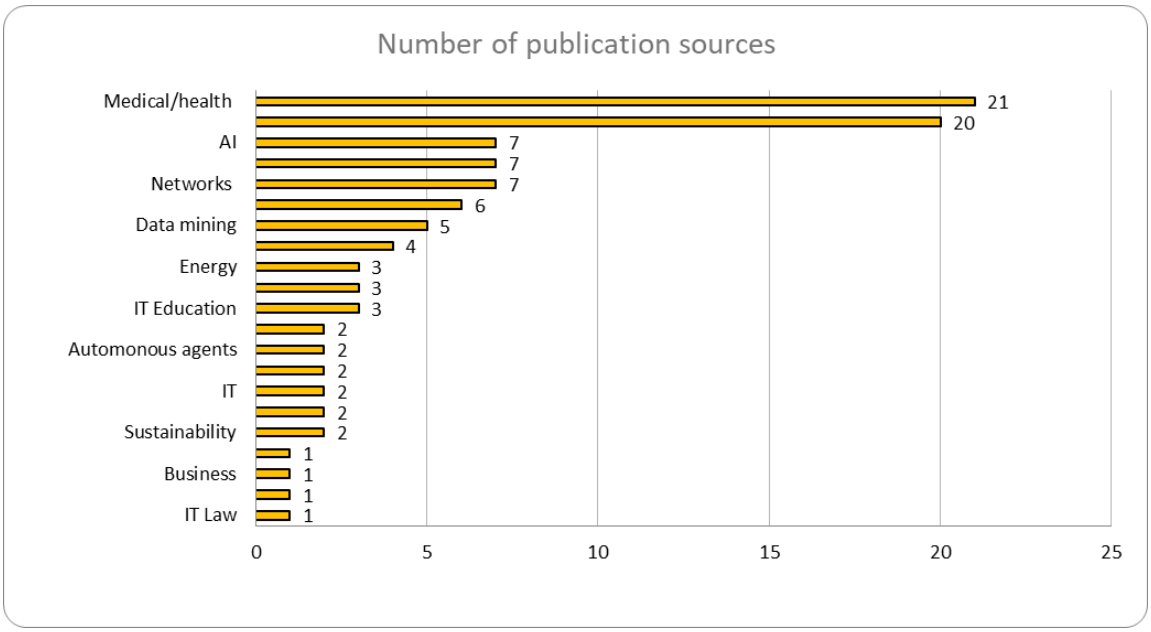

Figure 4. Human-centred AI publication sources by discipline

4. Conclusion

This paper aims to provide a metadata analysis of AI publications that have addressed the key principles of trustworthiness, safety, and reliability. The term and co-occurrence analysis reported here provides a range of themes and topics that are of particular significance that AI researchers should take into account as they conceptualize, design, and develop intelligent systems and services. The textual analysis of over 100 publications has demonstrated evidence of particular emphasis on the human factor and the usability and user acceptance of intelligent technologies. As health and medical technologies and car manufacturers are increasingly using AI agents and systems, the important principles of trustworthiness, safety, and reliability become critical components of not only the discourse of AI research and development, but also of the practical and pragmatic aspects of designing and developing responsible, explainable, and human-centred AI systems. Furthermore, the analysis provides insight into the themes and topics that future AI research and development should take into account in order to be human-centred, including such key topics as human control, usability analysis, trust management, user agent interaction, user robot interaction, user computer interaction along with such ethical principles as privacy, quality assurance, and predictive risk assessment. Future responsible and socially accountable AI-enabled systems can only be imagined, designed, and implemented if such foundational principles as trustworthiness, safety, and reliability are considered as integral constructs of any new and innovative AI system.

References

Carriço, G. (2018). The EU and artificial intelligence: A human-centred perspective. *European View*, *17*(1), 29-36.

Hagras, H. (2018). Toward human-understandable, explainable AI. *Computer*, *51*(9), 28-36.

He, H., Gray, J., Cangelosi, A., Meng, Q., McGinnity, T. M., & Mehnen, J. (2021). The challenges and opportunities of human-centred AI for trustworthy robots and autonomous systems. *IEEE Transactions on Cognitive and Developmental Systems*.

Shneiderman, B. (2022). *Human-Centered AI*. Oxford University Press.

Stanford University Human-centered AI institute: https://hai.stanford.edu/about

van Berkel, N., Tag, B., Goncalves, J., & Hosio, S. (2022). Human-centred artificial intelligence: a contextual morality perspective. *Behaviour & Information Technology*, *41*(3), 502-518.