EasyChair Preprint

№ 10833

Defining Human-Centered AI: a Comprehensive Review of HCAI Literature

Stefan Schmager, Ilias Pappas and Polyxeni Vassilakopoulou

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

September 5, 2023

DEFINING HUMAN-CENTERED AI: A COMPREHENSIVE REVIEW OF HCAI LITERATURE

Research full-length paper

Schmager, Stefan, University of Agder, Kristiansand, Norway, stefan.schmager@uia.no

Pappas, Ilias, University of Agder, Kristiansand, Norway, ilias.pappas@uia.no

Vassilakopoulou, Polyxeni, University of Agder, Kristiansand, Norway, polyxenv@uia.no

Abstract

This paper investigates the evolution of Human-Centered Artificial Intelligence (HCAI) as an emergent perspective on the design, development, and deployment of Artificial Intelligence (AI). It provides an overview of HCAI definitions, from the most established to the less common definitions found in the literature, highlighting the variety of emphases as well as the shared understandings among them. Based on the review, the paper proposes a new comprehensive HCAI definition, synthesizing the main features of the different definitions. Our HCAI definition highlights the necessity to understand the involved and affected people. To identify and understand their needs and values, the new definition highlights the use of Human-Centered Design methods. In an HCAI context, needs and values are mainly manifested through the concepts of Augmentation, and Control. Augmentation refers to the idea of using AI to enhance human capabilities and performance, rather than replacing human beings with machines. Control, on the other hand, deals with the governance and management of AI systems to ensure that they operate ethically and safely. The paper highlights the importance of collaboration between AI and IS researchers to advance the HCAI agenda and ensure that AI serves the interests of society.

Keywords: Human-Centered AI, HCAI, Artificial Intelligence, Human-Centered Design, Augmentation, Control.

1 Introduction

Day by day we encounter an abundance of news about novel AI technologies, breakthroughs, and scary stories, both from popular media as well as scientific research. There is no shortage of alarming wake-up calls, reminding us to pay close attention to how these technologies will evolve and to act accordingly. Correspondingly, there is a growing number of practitioners and researchers addressing questions on how to mitigate risks and make AI systems align with human needs and values (Dignum, 2019; Google, 2019; IBM, 2020; Microsoft, 2020; Schmager, 2022; Vassilakopoulou et al., 2022). It is still common to design and develop AI systems with the primary goal of creating algorithms that excel at performing specific tasks, e.g., image recognition, natural language processing, or autonomous driving. The current emphasis lies on optimizing performance metrics, such as accuracy, speed, or resource efficiency, rather than explicitly considering human values or societal impacts. As a consequence of these prevalent practices, Human-Centered Artificial Intelligence (HCAI) emerged as a different point of view on Artificial Intelligence (AI) design, development, and deployment that prioritizes human needs and aspirations. HCAI acknowledges the impact of AI systems on individuals, societies, and the overall human experience and puts humans at the center and its research strategies emphasize that the next frontier of AI is not just technological but also humanistic and ethical.

HCAI is a crucial perspective for the responsible design, development, and deployment of AI. Digital technologies may have a dual role, sometimes being part of the problem or facilitating solutions to existing problems (Dwivedi et al., 2022; Pappas et al., 2023). Placing human beings at the center allows the creation of AI systems that are more inclusive, trustworthy, and aligned with human values and goals (Schoenherr et al., 2023; Shneiderman, 2020a). It can guide AI design ensuring that AI can support and augment human abilities and find ways to address ethical implications and unintended consequences of AI (Xu, 2019). Yvonne Rogers calls HCAI “the new zeitgeist” (2022).

Different researchers from various disciplines have attempted to formulate their perspectives on HCAI introducing different definitions. However, a widely agreed-upon definition of HCAI has not yet been reached (Renz & Vladova, 2021). Having a shared and comprehensive definition as a conceptual bedrock could allow for clear and unambiguous communication and collaboration. It can help to avoid vague or ambiguous language, reducing the potential for misunderstandings, and enabling the alignment of strategies, actions, and common goals. It could promote consistency and coherence in discussions, decision-making, and problem-solving. Furthermore, a shared definition encourages critical thinking, as it provides a starting point for deeper exploration of the involved concepts, evaluating implications, weaknesses, and strengths. Overall, a shared definition facilitates a meaningful debate and will contribute to advancing the scientific discourse about the responsible introduction of AI technologies. Against this backdrop of ambiguity, this literature review aims to answer the research question: *How is Human-Centered AI defined in the existing literature?*

The objective of this work is to trace the evolution of HCAI mapping the ever-growing landscape of HCAI definitions in the literature and providing conceptual clarity by suggesting a comprehensive definition. This paper aims to accelerate research on HCAI, helping to produce AI-infused products, systems, and services with widespread benefits for individual users and the whole of society, including education, healthcare, environmental preservation, and community safety (Shneiderman, 2020b). The rest of the paper is structured as follows. First, the research method is presented. Then, different HCAI definitions are presented and synthesized into a new comprehensive definition. After that, a discussion is provided before concluding the paper.

2 Research Method

For this systematic literature review we applied the methodological framework by Kitchenham (2004), following her structured literature review process. The three steps within the framework consist of planning-, conducting-, and reporting the review. In the first step, we developed a detailed search protocol, defining specific search terms as well as inclusion/exclusion criteria. In the second step, the review was conducted. This includes identification, selection, appraisal of quality, evaluation, and synthesis of the literature. In the last step, the findings of the literature review are summarized and reported.

For this literature review, we conducted a database search in the SCOPUS research database on July 7th, 2022. The database has been chosen for being one of the most comprehensive databases of scientific literature and for its advanced search capabilities. In addition, SCOPUS employs rigorous quality control measures to ensure the quality and accuracy of the indexed literature, which helps to minimize the risk of low-quality or irrelevant articles. To collect resources as widely as possible, the defined search string was deliberately kept broad. An automated search with the search string TITLE-ABS-KEY ("human cent* AI" OR "human cent* artificial intelligence") has been conducted. By this, we ensured the search did consider American English as well as British English spellings of the search terms. The search was not limited by a time frame, since it was assumed that due to the novelty of the concept, a time limitation is not necessary. To ensure a high degree of relevance in the literature review corpus, the following exclusion criteria have been defined before the initial search and screening phases:

- Topic overviews not related to a conceptual understanding of AI.
- Studies discussing purely technical improvements.
- No AI relation, or AI only as an auxiliary aspect of the research.

Stage	**Description**	**Number**
Identification	Initial Results	215
1st Screening	After Abstract read	120
2nd Screening	After Full text read	109

Table 1. Review stages with the total number of sources at each stage

In the first screening stage, all abstracts from the initial list of 215 sources were read, which eliminated 95 sources as they matched the exclusion criteria. In the second screening phase, the remaining 120 sources have been fully read and assessed according to their suitability for the literature review. A total of 109 eligible, non-duplicate documents related to HCAI were identified.

The analysis was performed on a SCOPUS database export in the form of a spreadsheet, including information about Authors, Title, Year, Source, Abstract, and Keywords. The analysis examined whether each paper includes a definition for HCAI and if yes, if it reuses a pre-existing definition of HCAI or if it introduces a new one. If existing definitions were used, the respective references were marked in the spreadsheet. This coding was performed for all the papers in the corpus analyzed. This led to the identification of patterns and groupings within the literature, identifying the most used definitions, various combinations of definitions, and common concepts within the different definitions as well as the discovery that a significant number of publications don’t use a definition at all.

3 Findings

Approximately two out of three papers reviewed did not include a definition of the term Human-Centered Artificial Intelligence at all. In the remaining literature, we identified different definitions, with authors and professionals construing their conceptions into various, maybe similar yet still diverse meanings. In the paragraphs that follow, we first present the HCAI definitions by Shneiderman (2020a, 2020c, 2020d) which are the most widely used. After that, the paper provides a comprehensive overview of other HCAI definitions found in the literature, highlighting their various emphases, and shared understandings. In the final subsection, we provide a comprehensive HCAI definition synthesizing the literature.

3.1 HCAI as a paradigm shifting approach – Shneiderman’s definitions

The most widely used definition for HCAI is the one developed by Ben Shneiderman, a seasoned scholar in the field of Human-computer interaction (HCI). This definition reads as: *“HCAI focuses on amplifying, augmenting, and enhancing human performance in ways that make systems reliable, safe, and trustworthy. These systems also support human self-efficacy, encourage creativity, clarify responsibility, and facilitate social participation”* (Shneiderman, 2020a). By following the progression of how the term HCAI is used and described in Shneiderman’s topical publications, we can observe an evolution from being a term used to describe a conceptual framework, towards becoming a name for a paradigm-shifting approach for the development of AI technologies. Although the term HCAI has been used in the literature already from 1999 (Garcia, 1999), Shneiderman mentions it for the first time in his article “Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy” (2020a). In the article, the term HCAI is used for a two-dimensional framework that aims to enable high levels of human control as well as high levels of automation. The framework breaks the prevailing assumption of inverse proportionality for these two dimensions. The article’s argument is that an increase in automation does not inevitably implicate a decrease in human control or vice versa. Instead, systems should support both control and automation in order to be reliable, safe, and trustworthy. Such systems will increase human performance while supporting human self-efficacy, mastery, creativity, and responsibility.

Shneiderman develops this understanding of HCAI further in his article “Human-Centered Artificial Intelligence: Three Fresh Ideas” (Shneiderman, 2020c). Besides the two-dimensional framework of automation and control, he calls for an overall shift in language, imagery, and metaphors. In his later work “Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems” (Shneiderman, 2020d) he suggests 15 recommendations borrowing from software engineering practices to create reliable, safe, and trustworthy HCAI by enabling designers to translate widely discussed ethical principles into professional practices in large organizations with clear schedules. From this paper, it becomes clear that for Shneiderman, HCAI is not just a two-dimensional conceptual framework anymore, but it expands into considerations about processes and outcomes. Shneiderman refines his understanding of HCAI further in the paper “Human-Centered AI: A New Synthesis” (Shneiderman, 2021b), where he states that building AI-driven technologies that serve human needs requires combining AI-based algorithms with human-centered design (HCD) thinking. The fundamental conviction is that the adoption of user-centered design methodologies will lead to HCAI systems that support human goals, activities, and values. By that, Shneiderman indicates that HCAI is not the sole responsibility of a single discipline. Designers, engineers, product managers, government agencies, evaluators, and educators need to include HCAI ways of thinking. The goal is to enable a human-centered future with technologies that amplify, augment, and enhance human abilities and enhance human performance.

Shneiderman's work has been the conceptual foundation for many studies that take an HCAI perspective. Costabile et al. (2022) build upon the work of Shneiderman to explore three different interaction strategies for HCAI. In their study, they aim to develop a new class of tools for the interactive exploration of complex datasets and iterative meaning-making activities for humans with different levels of expertise. These tools can amplify, augment, and enhance human performance, in ways that make systems reliable, safe, and trustworthy. Vassilakopoulou and Pappas (2022), in their study on Chatbot – Human Agent handovers, draw from Shneiderman's work and define HCAI as the emerging discipline for AI-enabled systems that amplify and augment human abilities while preserving human control and ensuring ethically aligned design. Komischke (2021) uses Shneiderman's framework of human control and automation in the design and development of two digital productivity and collaboration applications use cases. Nagitta et al. (2022) examine the role of public procurement and procurement professionals in relation to HCAI principles and practical recommendations from Shneiderman (2020a, 2020d), highlighting the significance of HCAI for the benefit and safety of the public. Beckert (2021) uses the work of Shneiderman in his analysis of the state of play of implementing Trustworthy AI.

3.2 HCAI definitions beyond Shneiderman

Beyond the work by Shneiderman, we also identified other HCAI definitions used in the reviewed literature. These include definitions by Xu (2019) and Xu et al. (2022), the Stanford Institute for Human-Centered Artificial Intelligence (HAI, 2021), Riedl (2019), Auernhammer (2020), Dignum & Dignum (2020), and Holzinger (2022a, 2022b).

The definition developed by Xu (2019) and Xu et al. (2022) reads as: [HAI] "*includes three main components: 1) ethically aligned design, which creates AI solutions that avoid discrimination, maintain fairness and justice, and do not replace humans; 2) technology that fully reflects human intelligence, which further enhances AI technology to reflect the depth characterized by human intelligence (more like human intelligence); and 3) human factors design to ensure that AI solutions are explainable, comprehensible, useful, and usable".* Xu (2019) contributes to Human-Centered AI by proposing a framework that combines three components: "Ethically Aligned Design", "Technology Enhancement" and "Human Factors Design", which focuses on the intersection of AI and HCI. The main aim in Xu´s work is to explore how the HCI community can contribute to delivering AI solutions that are explainable, comprehensible, useful, and usable. This framework has been later refined showing that the individual components of Human Factors, Technology, and Ethics need to create synergies (Xu et al., 2022). The "Human Factors" component aims to ensure that AI solutions are comprehensible, useful, and usable to support human-driven decision-making processes. The "Technology" component is about defining human needs, designing, prototyping, and testing solutions together with users. This can contribute to developing human-controlled AI and to augmenting human abilities rather than replacing humans. The "Ethics" component relates to the creation of AI solutions that guarantee fairness, justice, and accountability. He et al. (2022) used Xu's framework in their study on challenges and opportunities for Trustworthy Robots and Autonomous Systems. They concluded that AI human-centeredness requires consideration of users and their cognition along with an understanding of reasoning processes and knowledge at the human level.

The Stanford Institute for Human-Centered Artificial Intelligence (HAI) states that Human-Centered AI aims *"[...] to augment the abilities of, address the societal needs of, and draw inspiration from human beings*" (HAI, 20121). The goal of HAI is to advance AI research, education, policy, and practice to improve the human condition, augment human intelligence, and thereby enhance human welfare by using machine intelligence. Stanford's HAI institute follows three objectives: technical reflection about the depth characterized by human intelligence; improving human capabilities rather than replacing them and focusing on AI's impact on humans (Stanford GDPi, 2018). In a New York Times article (2018), HAI Co-Director Fei-Fei Li states an aim to extend the popularity of human-centered approaches to AI toward more collaborative possibilities of mixed initiatives between human workers and AI agents. Li gives an example of how AI automation should focus on enhancing the strengths of humans "like dexterity and adaptability" by "keeping tabs on more mundane tasks and protecting against human error, fatigue, and distraction" (Wang et al., 2019).

Riedl (2019) proposed the following HCAI definition: *"Human-centered AI is a perspective on AI and ML [machine learning] that intelligent systems must be designed with awareness that they are part of a larger system consisting of human stakeholders, such as users, operators, clients, and other people in close proximity".* This means, an understanding of human sociocultural norms as part of a theory of mind as well as capabilities to produce explanations that nonexpert end-users can understand, are needed. For Riedl, HCAI means building systems to understand the often culturally specific expectations and needs of humans and to help humans understand the systems in return. Riedl breaks human-centered AI into two critical capacities, understanding humans, and being able to help humans understand AI. Riedl´s work has served as the conceptual foundation for the study by Elahi et al. (2021) on improving the privacy of older app users in smart cities. Also, Böckle et al. (2021) used the HCAI definition by Riedl to guide the design of their study on the effect of personality traits on trust in AI-enabled user interfaces.

According to Auernhammer's (2020) definition, *"Human-centered AI needs to focus on three integrated perspectives when designing AI systems: rationalistic (technology), humanistic (people), and judicial (policies)."* Auernhammer argues that pan-disciplinary research from fields like psychology, cognitive science, computer science, engineering, business management, law, and design is required to develop a genuinely human-centered approach for AI, since in essence, HCAI is about people. The work by Auernhammer has been used by Subramonyam et al. (2021) for the development of a Process Model for Co-Creating AI Experiences (AIX). Subramonyam and colleagues provide designers with practical guidance on how to work with AI systems as a design material and offer design considerations for incorporating data probes.

Dignum & Dignum (2020) describe an AI system as "Human-centered" if the system does not operate in isolation but is socially aware of performing its tasks for someone, within a local and temporal context. It is argued that AI systems are socio-technical systems in the sense that the social context of how these systems are developed, used, and acted upon is a fundamental consideration. This means, that for a Human-Centered approach to AI, the technical component cannot be separated from the socio-technical system (Dignum, 2019; Schoenherr et al., 2023). A perspective that is shared with Riedl (2018). The High-Level Expert Group of the European Commission (AI-HLEG, 2019) where Dignum takes part, developed the AI-HLEG-AI guidelines that include human-centricity. Although the group states that their ultimate ambition is to reach trustworthy AI, the formulated guidelines provide a definition for HCAI. They define a human-centric approach to AI as one in which "*humans enjoy a unique and inalienable moral status of primacy in the civil, political, economic, and social fields. AI systems need to be human-centric, resting on a commitment to their use in the service of humanity and the common good, intending to improve human welfare and freedom".*

Holzinger (2022a, 2022b) defines HCAI as a synergistic approach of "artificial intelligence" and "natural intelligence" to empower, amplify, and augment human performance, rather than replace people. Its goal is to promote the robustness of AI algorithms and to align AI solutions with human values, ethical principles, and legal requirements to ensure safety and security, enabling trustworthy AI. Steels (2020) argues that human-centric AI is only going to be possible when AI comes to grips with meaning and understanding. They are building upon the work by Nowak et al. (2018) which points to HCAI as a "possible path" to avoid dystopian developments. The authors distinguish the way AI is being built into "Function-Oriented AI" and "Human-Centered AI". HCAI is envisioned as synergistically working together with humans for the benefit of humans and human society, focusing on enhancing and empowering humans rather than replacing and controlling them.

Several articles combine more than one definition of HCAI. For instance, Herrmann (2022) employs the HCAI definitions of Shneiderman (2020d) and the framework by Xu (2019) in research on interaction modes for promoting human capabilities. The identified interaction modes highlight both human and AI strengths. Examples include the provision of explanations and possibilities for exploration, testing, and re-training with human involvement and keeping humans in control by allowing for intervention and vetoing. Another example of a combination of definitions is the research by Yang et al. (2021). Yang and colleagues in their conceptual work on smart learning environments state that HCAI can be interpreted from two perspectives. The first is AI under human control, describing the interplay between human control and AI automation (Shneiderman, 2020a). The other is AI on the human condition, which refers to having explainable and interpretable computation and judgment processes and continuous adjustments of AI to societal phenomena (HAI, 2021).

3.3 A comprehensive HCAI definition based on the literature

The table that follows provides an overview of the most used definitions of Human-Centered AI identified within the reviewed literature (Table 2).

Definition	**Source**
HCAI focuses on amplifying, augmenting, and enhancing human performance in ways that make systems reliable, safe, and trustworthy. These systems also support human self-efficacy, encourage creativity, clarify responsibility, and facilitate social participation	Shneiderman (2020a)
[HAI] *includes three main components: 1) ethically aligned design, which creates AI solutions that avoid discrimination, maintain fairness and justice, and do not replace humans; 2) technology that fully reflects human intelligence, which further enhances AI technology to reflect the depth characterized by human intelligence (more like human intelligence); and 3) human factors design to ensure that AI solutions are explainable, comprehensible, useful, and usable.*	Xu (2019)
[Human-Centered AI aims] to augment the abilities of, address the societal needs of, and draw inspiration from human beings.	HAI (2021)
Human-centered AI is a perspective on AI and ML that intelligent systems must be designed with awareness that they are part of a larger system consisting of human stakeholders, such as users, operators, clients, and other people in close proximity.	Riedl (2019)
Human-centered AI needs to focus on three integrated perspectives when designing AI systems: rationalistic (technology), humanistic (people), and judicial (policies).	Auernhammer (2020)
Human-centered means that a system should have the human partner always as part of the focus for deliberation. This means that any task of the AI system should not be done in isolation, but the task should be done for someone, in some context (place and time). And if the actions of the AI system affect people directly or indirectly it should be aware of this and take it into consideration when deliberating.	Dignum & Dignum (2020)
AI systems need to be human-centric, resting on a commitment to their use in the service of humanity and the common good, intending to improve human welfare and freedom.	AI-HLEG (2019)
Human-centered AI we define as a synergistic approach to align AI solutions with human values, ethical principles, and legal requirements to ensure safety and security, enabling trustworthy AI.	Holzinger (2022a)
By this [HCAI] we mean designing AI systems that enhance human capacities and improve human experiences rather than replacing them through automation.	Rogers (2019)

Table 2. Overview of Human-Centered AI definitions in the literature

The literature review revealed much conceptual overlap among the identified definitions coming from the different scholars of HCAI. At the same time, the review also highlights the diversity in emphases and approaches toward an understanding of what Human-Centered Artificial Intelligence could entail.

Based on the different definitions, we are proposing a new comprehensive definition of HCAI to represent the richness of the scholarly understandings:

> *Human-Centered AI (HCAI) focuses on understanding purposes, human values and desired AI properties in the creation of AI systems by applying Human-Centered Design practices. HCAI seeks to augment human capabilities while maintaining human control over AI systems, by considering the necessity, context, and ethical and legal conditions of the AI system as well as promoting individual and societal well-being."*

Our definition aims to emphasize the fundamentally humane character of HCAI while also encompassing its contributing constituents. By incorporating Human-Centered Design methodologies, e.g., stakeholder participation, HCAI underscores the constant reflection of whether an envisioned AI system is in accordance with the pluralism of human needs and values. Further, our definition highlights context sensitivity, including the acknowledgment of stakeholder diversity, a comprehension of the context of use, and the awareness that an AI system is not a single entity, but rather a part of a larger structure. Understanding the characteristics of an AI system, including scope, usage implications, and sociocultural context are crucial factors of HCAI. Finally, our definition addresses the consideration of ethical and legal requirements, to ensure a responsible and lawful design, development, and deployment of an AI system. In essence, our definition delineates the overarching objective of HCAI to consider and promote the well-being of individuals as well as the whole of society.

4 Discussion

This literature review illustrates multiple takes on what "Human-Centered Artificial Intelligence" could mean. This is not surprising, since defining a term that is linked to a constantly evolving technology like AI, is like trying to hit a moving target. Deconstructing the term HCAI into its two parts "Human-Centeredness - HC" and "Artificial Intelligence - AI" further illustrates this difficulty. While there are definitions available for human-centeredness, for example, from HCI, Interaction-, and UX Design (Xu, 2019), a universally agreed definition of AI is yet to be found. And even if such a lack of consensus is accepted, the question remains if HCAI just describes the intersection of HC and AI, or if it constitutes something greater than the sum of its parts. As the literature review unveiled, for some, HCAI is understood as the amalgamation of Human-Centered Design and AI. Several definitions highlight the necessity of incorporating Human-Centered Design methods in the design and development processes of AI systems. Yet for others, HCAI constitutes nothing less than a paradigm shift, moving beyond the prevalent technology-centered approaches towards AI driven by human values.

Developing a common and shared definition can play an important role in advancing scientific research by promoting clarity, collaboration, and progress. In the realm of scientific inquiry having a shared understanding of key concepts and terms is essential to foster clear communication among researchers, minimizing misunderstandings. A shared understanding promotes a meaningful exchange of ideas that allows scholars and practitioners to build upon each other's work, develop new hypotheses, and advance the scientific discourse. A common conceptual ground can encourage collaboration among researchers as well as with practitioners and help to align efforts, combine expertise, and work towards common goals. Furthermore, such a shared understanding can enhance the reliability and reproducibility of research findings. This is crucial for validating and building upon existing research, strengthening the knowledge base, and fostering knowledge transfer within the scientific community. At the same time agreed-upon definitions and a shared understanding of involved concepts facilitate critical thinking, fostering intellectual growth and driving scientific progress. This allows for focused debates, evaluating the strengths and weaknesses of different approaches, and critically analyzing the implications of research outcomes.

Our analysis of existing HCAI definitions identified a common understanding that a human-centered approach to AI foregrounds human needs and values. This is most notably manifested in the two concepts Augmentation and Control. The maxim of "Augmentation instead of replacement" is based on the understanding that technology is created with the purpose of supporting humans, not making them redundant. Augmentation is ingrained in HCAI conceptualizations in different ways. For Shneiderman, HCAI is about the creation of super tools, powered by advanced technologies like deep neural networks but still considered tools because they come into existence to support their users (Shneiderman, 2020a, 2020c, 2020d). Xu et al. (2022) have included the postulation of not replacing humans in the "Ethics" component of their model for the human-centered development of AI. Xu and colleagues shift the perspective from a purely technical question, i.e., "Can we?" towards an ethical one, i.e., "Should we?". Similarly, in the vision of Stanford's Human-centered AI Institute, the improvement of human capabilities rather than their replacement is one of three core objectives (HAI, 2021). The aim for human augmentation is also evident in the synergistic approaches to HCAI by Holzinger (2022a) and Nowak et al. (2018). These papers share the notions of empowering, amplifying, and augmenting human performance, rather than replacing people.

Furthermore, the concept of control is also closely connected to HCAI in the literature. Shneiderman argues that control and automation are not necessarily two ends of the same spectrum, but rather, two separate dimensions. In his framework, high levels of control and high levels of automation are not mutually exclusive. Shneiderman claims that both control and automation are needed for HCAI systems (Shneiderman, 2020a). Xu et al. (2022) highlight a shift from human-centered automation to human-controlled autonomy. The same understanding is implied in the definitions by Holzinger (2022a) and in Xu's earlier work (2019). The concept of control raises questions around the ultimate power of decision, considering how human-beings are involved in decision making processes when AI is also involved.

To gauge appropriate levels of augmentation and control, our definition highlights the importance of established Human-Centered Design (HCD) methods and practices. HCD describes a creative approach to problem-solving that starts with understanding the people involved and designing around their needs and values. An HCD approach is described as cultivating deep empathy with the people you're designing with, generating ideas, building different prototypes, sharing what you've made together, and eventually putting your innovative new solution out in the world (IDEO, 2023). The US Office of Science and Technology Policy in its Strategic Plan on National AI Research and Development, has recently explicitly favored human factors, usability, and human-centered design research methods (OSTP, 2023). In particular, the report argues for the analysis of user needs and requirements through iterative design methods to understand and address the ethical, legal, and societal implications of AI and to ensure safety and security.

Enhancing human abilities with the help of technology, while exploring appropriate levels of automation, supervision, and decision-making are known objects of inquiry. Back in 1989, Banon and Schmidt noted that by changing the allocation of functions between humans and their implements, changes in technology induce changes in work organization (Banon & Schmidt, 1989). As AI becomes widespread and ubiquitous across work settings, and more functions get delegated to AI-infused systems, the relevance of HCAI becomes clear. Liikkanen (2019) describes that human-centered design will be crucial in further defending humans, particularly underprivileged users at risk of being mistreated by AI.

5 Conclusion

This literature review provides an overview of HCAI definitions, from the most established to the less common ones. It highlights the partly shared conceptual understanding, but also the existing diversity of emphases among them. Based on the review, we are proposing a new comprehensive HCAI definition, synthesizing the main attributes of the different existing definitions. Our proposed HCAI definition highlights the necessity to engage with and understand the involved and affected people. To identify and understand their needs and values, our new definition also highlights the use of HCD methods. In regard to such needs and values, a particular focus has been identified for the concepts of Augmentation, and Control. Augmentation describes the idea of enhancing human capabilities and performance using AI, rather than replacing human beings with machines. The concept of control deals with aspects of governance and management of AI systems to ensure they operate ethically and safely.

Overall, the variety of HCAI definitions indicates a steadily growing interest which gives an optimistic outlook for the future. According to Rogers (2021), we are currently reimagining rather than revisiting longstanding dystopian visions of AI. She describes the nascent HCAI research as an eclectic discipline full of inclusive voices, doing exciting, enabling, and empowering work. The comprehensive definition introduced can be used as a foundation for researchers and practitioners to ensure a common understanding of the concept enabling consistency, communication, and collaboration.

Analyzing the landscape of definitions for an emerging and constantly evolving concept doesn't come without limitations. The first limitation is of a rather practical nature, as the wealth of literature related to Human-Centered AI is rapidly increasing. The pace of new academic output for this highly relevant topic is only exceeded by the number of technological breakthroughs it tries to examine. Furthermore, there might be literature that describes the same fundamental idea of HCAI, which has not been captured by our keyword search if it uses different terminologies. Another limitation stems from criticism towards the general HCD idea. Norman (2005) states that HCD has become such a dominant theme, that its principles can be misleading, wrong, or at times even harmful. A more evolutionary criticism of HCD has been formulated by scholars suggesting "More-Than-Human design" which extends the universe of design beyond human needs and values (Giaccardi & Redström, 2020; Nicenboim et al., 2020; Coskun et al., 2022).

Human involvement in the creation and critique of the design of AI technologies demonstrates how society can benefit from having many kinds of human-machine interaction at its fingertips rather than focusing on the consequences of a seismic shift in machine autonomy. Going forward, the field of AI will have far-reaching impacts within the workplace and beyond. As wonderfully phrased by Yang et a. (2021), "AI may be a current trend, but humanistic beauty is eternal".

References

Auernhammer, J. (2020). *Human-centered AI: The role of Human-centered Design Research in the development of AI.* In Boess, S., Cheung, M. and Cain, R. (eds.), Synergy - DRS International Conference 2020, 11-14 August, Held online. https://doi.org/10.21606/drs.2020.282

Bannon, L. J., & Schmidt, K. (1989). *CSCW: Four characters in search of a context.* In ECSCW 1989: Proceedings of the First European Conference on Computer Supported Cooperative Work. Computer Sciences Company, London.

Beckert, B. (2021, September). *The European way of doing Artificial Intelligence: The state of play implementing Trustworthy AI.* In 2021 60th FITCE Communication Days

Böckle, M., Yeboah-Antwi, K., & Kouris, I. (2021). *Can you trust the black box? The effect of personality traits on trust in AI-enabled user interfaces.* In Artificial Intelligence in HCI: Second International Conference, AI-HCI 2021, Held as Part of the 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings (pp. 3-20). Cham: Springer International Publishing.

Coskun, A., Cila, N., Nicenboim, I., Frauenberger, C., Wakkary, R., Hassenzahl, M., ... & Forlano, L. (2022). *More-than-human Concepts, Methodologies, and Practices in HCI.* In CHI Conference on Human Factors in Computing Systems Extended Abstracts (pp. 1-5).

Costabile, M. F., Desolda, G., Dimauro, G., Lanzilotti, R., Loiacono, D., & Matera, M and Zancanaro, M. (2022). *A Human-centric AI-driven Framework for Exploring Large and Complex Datasets.* Proceedings of CoPDA2022 - Sixth International Workshop on Cultures of Participation in the Digital Age: AI for Humans or Humans for AI? June 7, 2022, Frascati (RM), Italy

Dignum, V. (2019). *Responsible artificial intelligence: how to develop and use AI in a responsible way.* Cham: Springer.

Dignum, F., & Dignum, V. (2020). *How to center AI on humans.* In NeHuAI 2020, 1st International Workshop on New Foundations for Human-Centered AI, Santiago de Compostela, Spain, September 4, 2020 (pp. 59-62).

Dwivedi, Y. K., Hughes, L., Kar, A. K., Baabdullah, A. M., Grover, P., Abbas, R., ... & Wade, M. (2022*). Climate change and COP26: Are digital technologies and information management part of the problem or the solution? An editorial reflection and call to action.* International Journal of Information Management, 63, 102456.

Elahi, H., Castiglione, A., Wang, G., & Geman, O. (2021). *A human-centered artificial intelligence approach for privacy protection of elderly App users in smart cities.* Neurocomputing, 444, (pp. 189-202).

Garcia, O. (1999). *An Approach to Complexity from a Human-Centered Artificial Intelligence Perspective.* In Encyclopedia of Computer Science and Technology, Vol. 40 (A. Kent and J. G. Williams, eds.), Marcel Dekker, New York (pp. 1-16).

Giaccardi, E., & Redström, J. (2020). *Technology and more-than-human design.* Design Issues, *36*(4), (pp. 33-44).

Google. (2019). *Responsible AI practices.* Retrieved 16 February from https://ai.google/responsibilities/responsible-ai-practices/

HAI Research. (2021). *Guiding Human-Centered AI.* Stanford Institute for Human-Centered Artificial Intelligence. Retrieved January 28, 2023, from https://hai.stanford.edu/research

He, H., Gray, J., Cangelosi, A., Meng, Q., McGinnity, T. M., & Mehnen, J. (2020). *The challenges and opportunities of artificial intelligence for trustworthy robots and autonomous systems.* In 2020 3rd International Conference on Intelligent Robotic and Control Engineering (IRCE) (pp. 68-74). IEEE.

Herrmann, T. (2022). *Promoting Human Competences by Appropriate Modes of Interaction for Human-Centered-AI.* In Artificial Intelligence in HCI: 3rd International Conference, AI-HCI 2022, Held as Part of the 24th HCI International Conference, HCII 2022, Virtual Event, June 26–July 1, 2022, Proceedings (pp. 35-50). Cham: Springer International Publishing.

High-Level Expert Group on Artificial Intelligence (AI-HLEG). (2019). *Ethics Guidelines for Trustworthy AI.* Brussels: European Commission. https://ec.europa.eu/futurium/en/ai-alliance-consultation/

Holzinger, A., Saranti, A., Angerschmid, A., Retzlaff, C. O., Gronauer, A., Pejakovic, V., Medel-Jimenez, F., Krexner, T., Gollob, C. & Stampfer, K. (2022a). *Digital transformation in smart farm and forest operations needs human-centered AI: challenges and future directions.* Sensors, 22(8), 3043.

Holzinger, A., Kargl, M., Kipperer, B., Regitnig, P., Plass, M., & Müller, H. (2022b). *Personas for artificial intelligence (AI) an open source toolbox.* IEEE Access, 10, 23732-23747.

IBM. (2020). *AI ethics (IBM's multidisciplinary, multidimensional approach helping advance responsible AI)*. Retrieved 16 February from https://www.ibm.com/artificial-intelligence/ethics

IDEO (2023). *Design thinking frequently asked questions (FAQ)*. https://designthinking.ideo.com/faq/whats-the-difference-between-human-centered-design-and-design-thinking. Retrieved at 29 Jun 2023.

Kitchenham, B. (2004). *Procedures for performing systematic reviews.* Keele, UK, Keele University, 33(2004), 1-26.

Komischke, T. (2021). *Human-centered artificial intelligence considerations and implementations: a case study from software product development.* In Artificial Intelligence in HCI: Second International Conference, AI-HCI 2021, Held as Part of the 23rd HCI International

Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings (pp. 260-268). Cham: Springer International Publishing.

Li, F. (2018). *Opinion | How to Make A.I. That's Good for People.* The New York Times. https://www.nytimes.com/2018/03/07/opinion/artificial-intelligence-human.html

Liikkanen, L. A. (2019). *It ain't nuttin' new – interaction design practice after the ai hype.* In Human-Computer Interaction–INTERACT 2019: 17th IFIP TC 13 International Conference, Paphos, Cyprus, September 2–6, 2019, Proceedings, Part IV 17 (pp. 600-604). Springer International Publishing.

Microsoft. (2020). *Responsible AI (policies, practices, and tools that make up a framework for Responsible AI by Design)*. Retrieved 16 February from https://www.microsoft.com/en-us/ai/responsible-ai

Nagitta, P. O., Mugurusi, G., Obicci, P. A., & Awuor, E. (2022). *Human-centered artificial intelligence for the public sector: The gate keeping role of the public procurement professional.* Procedia Computer Science, 200, (pp. 1084-1092).

US Office of Science and Technology Policy (OSTP) - Select Committee on Artificial Intelligence. (2023). *National Artificial Intelligence Research and Development Strategic Plan 2023.* (https://digital.library.unt.edu/ark:/67531/metadc2114122/: accessed June 29, 2023), University of North Texas Libraries, UNT Digital Library, https://digital.library.unt.edu.

Nicenboim, I., Giaccardi, E., Søndergaard, M. L. J., Reddy, A. V., Strengers, Y., Pierce, J., & Redström, J. (2020). *More-than-human design and AI: in conversation with agents.* In Companion publication of the 2020 ACM designing interactive systems conference (pp. 397-400).

Norman, D. A. (2005). *Human-centered design considered harmful.* Interactions, 12(4), 14-19.

Nowak, A., Lukowicz, P., & Horodecki, P. (2018). *Assessing artificial intelligence for humanity: Will AI be the our biggest ever advance? Or the biggest threat* [Opinion]. IEEE Technology and Society Magazine, 37(4), (pp. 26-34).

Pappas, I. O., Mikalef, P., Dwivedi, Y. K., Jaccheri, L., & Krogstie, J. (2023). *Responsible Digital Transformation for a Sustainable Society.* Information Systems Frontiers, 1-9.

Renz, A., & Vladova, G. (2021). *Reinvigorating the discourse on human-centered artificial intelligence in educational technologies.* Technology Innovation Management Review, 11(5).

Riedl, M. O. (2019). *Human-centered artificial intelligence and machine learning.* Human Behavior and Emerging Technologies, 1(1), 33-36.

Rogers, Y., Brereton, M., Dourish, P., Forlizzi, J., & Olivier, P. (2021). *The dark side of interaction design.* Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, Article 152, 1–2. https://doi.org/10.1145/3411763.3450397

Rogers, Y. (2022). *Commentary: human-centred AI: the new zeitgeist.* Human–Computer Interaction, 37(3), (pp. 254-255).

Schmager, S. (2022). *From commercial agreements to the social contract: human-centered AI guidelines for public services.* Proceedings of the 14th Mediterranean Conference on Information Systems (MCIS 2022). Association for Information Systems (AIS).

Schoenherr, J. R., Abbas, R., Michael, K., Rivas, P., & Anderson, T. D. (2023). *Designing AI using a human-centered approach: Explainability and accuracy toward trustworthiness.* IEEE Transactions on Technology and Society, 4(1), 9-23.

Shneiderman, B. (23 Mar 2020a). *Human-centered artificial intelligence: Reliable, safe & trustworthy.* International Journal of Human–Computer Interaction, 36(6), (pp. 495-504).

Shneiderman, B. (2020b). *Design lessons from AI's two grand goals: human emulation and useful applications.* IEEE Transactions on Technology and Society, 1(2), 73-82.

Shneiderman, B. (2020c). *Human-centered artificial intelligence: Three fresh ideas.* AIS Transactions on Human-Computer Interaction, 12(3), (pp. 109-124).

Shneiderman, B. (2020d). *Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems.* ACM Transactions on Interactive Intelligent Systems (TiiS), 10(4), (pp. 1-31).

Shneiderman, B. (2021a). *Tutorial: Human-centered AI: Reliable, safe and trustworthy.* In 26th International Conference on Intelligent User Interfaces-Companion (pp. 7-8).

Shneiderman, B. (2021b). *Human-centered AI: A new synthesis.* In Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part I 18 (pp. 3-8). Springer International Publishing.

Stanford GDPi (2018). *Human-Centered AI: Building Trust, Democracy and Human Rights by Design.* Medium. https://medium.com/stanfords-gdpi/human-centered-ai-building-trust-democracy-and-human-rights-by-design-2fc14a0b48af

Steels, L. (2020). *Personal dynamic memories are necessary to deal with meaning and understanding in human-centric AI.* In NeHuAI@ ECAI (pp. 11-16).

Subramonyam, H., Seifert, C., & Adar, E. (2021). *Towards a process model for co-creating AI experiences.* In Designing Interactive Systems Conference 2021 (pp. 1529-1543).

Vassilakopoulou, P., & Pappas, I. O. (2022). *AI/Human augmentation: a study on chatbot–human agent handovers.* In Co-creating for Context in the Transfer and Diffusion of IT: IFIP WG 8.6 International Working Conference on Transfer and Diffusion of IT, TDIT 2022, Maynooth, Ireland, June 15–16, 2022, Proceedings (pp. 118-123). Cham: Springer International Publishing.

Vassilakopoulou, P., Parmiggiani, E., Shollo, A., & Grisot, M. (2022). *Responsible AI: Concepts, critical perspectives and an Information Systems research agenda.* Scandinavian Journal of Information Systems, 34(2), 3.

Wang, D., Weisz, J. D., Muller, M., Ram, P., Geyer, W., Dugan, C., Tausczik, Y., Samulowitz, H. & Gray, A. (2019). *Human-AI collaboration in data science: Exploring data scientists' perceptions of automated AI.* Proceedings of the ACM on human-computer interaction, 3(CSCW), (pp. 1-24).

Xu, W. (2019). *Toward human-centered AI: a perspective from human-computer interaction.* interactions, 26(4), (pp. 42-46).

Xu, W., Dainoff, M. J., Ge, L., & Gao, Z. (2022). *Transitioning to human interaction with AI systems: New challenges and opportunities for HCI professionals to enable human-centered AI.* International Journal of Human–Computer Interaction, 39(3), (pp. 494-518).

Yang, S. J., Ogata, H., Matsui, T., & Chen, N. S. (2021). *Human-centered artificial intelligence in education: Seeing the invisible through the visible.* Computers and Education: Artificial Intelligence, 2, 100008.