Understandings and perspectives of human-centered AI – a transdisciplinary literature review

Uta WILKENS[1], Christian COST REYES[1], Tim TREUDE[1], Annette KLUGE[2]

[1] Institut für Arbeitswissenschaft, Ruhr-Universität Bochum
Universitätsstraße 150, D-44801 Bochum
[2] Lehrstuhl Arbeits-, Organisations- & Wirtschaftspsychologie
Ruhr-Universität Bochum
Universitätsstraße 150, D-44801 Bochum

Abstract. The paper presents findings from a systematic literature review exploring the meaning of human-centered AI. The review includes 85 papers from certain disciplines and leads to a distinction of five co-existing perspectives: (1) a deficit-oriented, (2) a data reliability-oriented, (3) a protection-oriented, (4) a potential-oriented and (5) a political-oriented understanding of how to reach human centricity while using AI in the workplace. Each perspective gives emphasis to another core dimension for evaluating the level of human-centricity. This goes from compensating individual weaknesses with the help of AI to enhancing data reliability and protecting individual integrity, to leveraging individual potential and guaranteeing individual control over AI. Each perspective is exemplified with a use case.

Keywords: artificial intelligence, human-centricity, human-in-the-loop, reliability, workplace

1. Introduction and aim of analysis

The notion of “human-centered” artificial intelligence (AI) gains high attention in current writings. This implies that the subject (human being) and the object (technology in terms of AI) might unfold a new interaction intensity as collaborating partners in a work system (Onnasch et al. 2016; Wilson & Daugherty 2018; Muhle 2019). At least the discourse aims at avoiding a technology-dominated focus as it considers the complementary potential between AI and the human being (Xu 2019; Wilkens 2020). So far, the notion seems rather to be used as an umbrella term as there is no common ground or widely shared understanding of what human-centricity exactly means. Different disciplines provide a range of interpretations which tend to go in hand with their implicit basic beliefs in human nature and human behavior at work (see McGregor 1960) respectively the potential and deficits of AI (e.g. Garcia-Magarino 2019).

In order to develop work systems where people and AI interact and perform tasks together there is a need for a theoretical foundation and empirical validation of the criteria for the human-centricity of AI. Wilkens et al. (2019) developed a classification which shows how AI is related to the individual or organizational learning process. This follows the idea that human-centricity indicates the level of recognition of individual intelligence. But there are many other implicit assumptions of what human-centricity of AI means. As these assumptions have impact on the development and use of AI in the workplace this paper aims at explicating the spectrum of co-existing interpretations of the human-centricity of AI. Understanding different basic beliefs is an important prerequisite to evaluate ongoing developments and to define criteria for a human-centered design of AI. The outline is based on a systematic literature review.

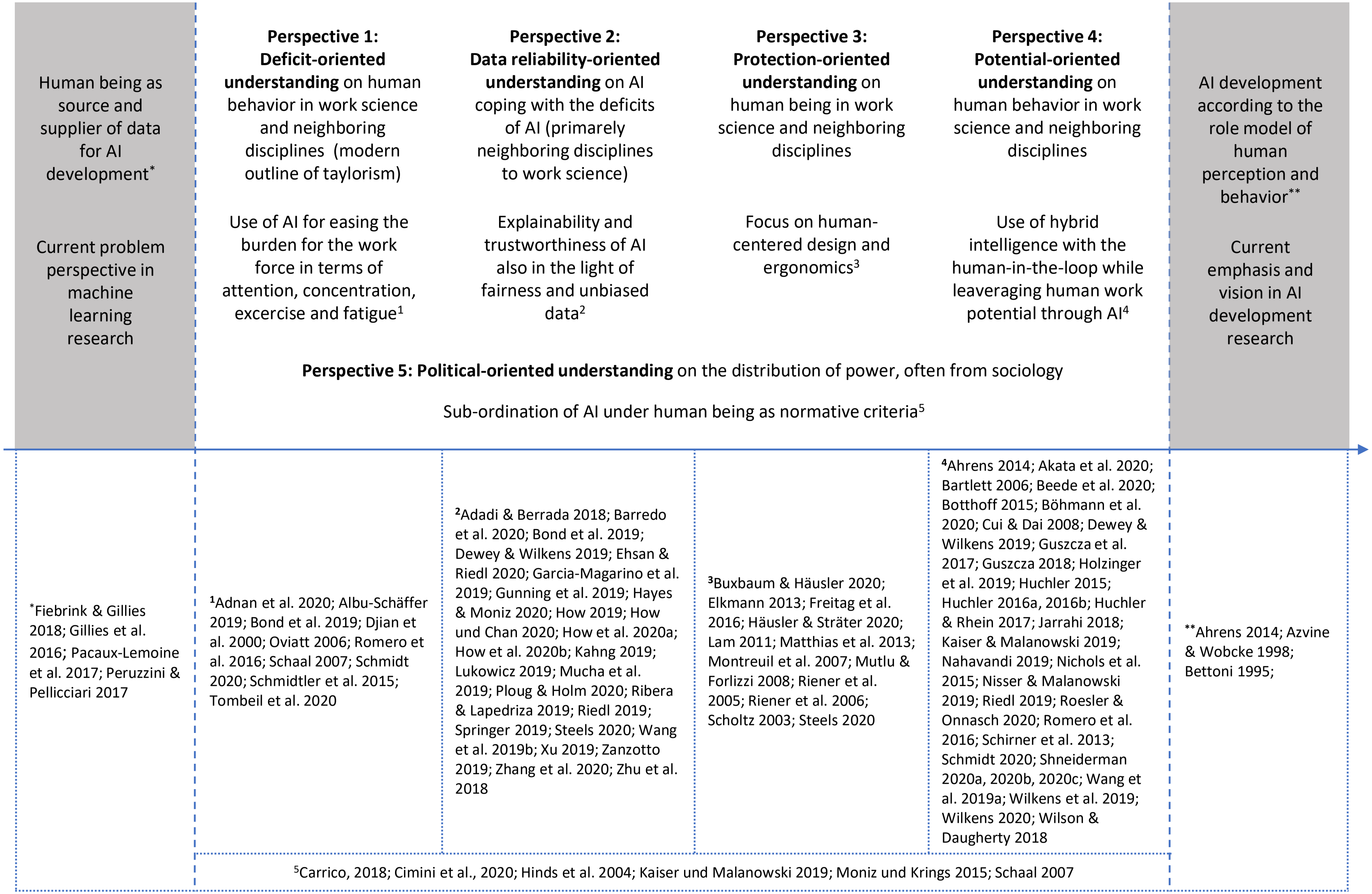
2. Methodology

The methodology in use is a systematic transdisciplinary literature review aligned to the method from Riasanow et al. (2019) including combinations of the keywords "human-centered" or "people-centered" with "artificial intelligence" or "AI" or "machine learning" or similar combinations of "human" and "robotics" – both in English and German language. The search includes journal publications, project reports, books and book chapters from engineering, information science, medicine, psychology, management, philosophy, human factor studies, work science, educational science etc. As many disciplines elaborate on the same subject but enter the field of AI from different perspectives and theoretical fundaments it is important to include this high variety in the literature review. We consciously avoided to focus on specific journals and rankings as this does not correspond equally with the publication strategy of different disciplines and as publications of high novelty are not necessarily published in highly ranked journals. In total we explored 133 publications and selected 85 for the further course of analysis. The selection process followed the idea that the source – based on the inherent understanding of human centricity - has the potential to directly or indirectly contribute to future socio-technical system design. We excluded publications with a pure focus on IT security, privacy, technology development, philosophy and politics as long as they did not consider AI related to labor, workplaces or job design.

The data evaluation of the literature review is based on a qualitative content analysis with an open coding process. Four authors participated in this process and developed a shared understanding of the different directions to be find in the selected papers. There was no ex-ante definition of codes; the authors' hermeneutical interpretation (Prasad 2002) of papers was in the center of the data analysis. As a result five different basic understandings of what human-centricity means and five related dimensions specifying different levels of human-centered AI could be explored. As far as possible the identified basic understandings are also related to the disciplines where they are primarily represented.

3. Findings

Many disciplines address issues of human-centered AI which means that the construct itself has to be treated as a transdisciplinary field of research. Disciplines that came up by our search strategy are: AI development research, Anthropology, Education Studies, Engineering and Industry Research, Human Factor Studies, Law, Logistics & Transport, Management Studies especially Finance and Marketing, Medicine, Military Studies, Philosophy & Ethics, Psychology, Sociology, Socionics & Politics, Work Science.



Human being as source and supplier of data for AI development*

Current problem perspective in machine learning research

Perspective 1: Deficit-oriented understanding on human behavior in work science and neighboring disciplines (modern outline of taylorism)

Use of AI for easing the burden for the work force in terms of attention, concentration, excercise and fatigue[1]

Perspective 2: Data reliability-oriented understanding on AI coping with the deficits of AI (primarely neighboring disciplines to work science)

Explainability and trustworthiness of AI also in the light of fairness and unbiased data[2]

Perspective 3: Protection-oriented understanding on human being in work science and neighboring disciplines

Focus on human-centered design and ergonomics[3]

Perspective 4: Potential-oriented understanding on human behavior in work science and neighboring disciplines

Use of hybrid intelligence with the human-in-the-loop while leaveraging human work potential through AI[4]

AI development according to the role model of human perception and behavior**

Current emphasis and vision in AI development research

Perspective 5: Political-oriented understanding on the distribution of power, often from sociology

Sub-ordination of AI under human being as normative criteria[5]

*Fiebrink & Gillies 2018; Gillies et al. 2016; Pacaux-Lemoine et al. 2017; Peruzzini & Pellicciari 2017

[1] Adnan et al. 2020; Albu-Schäffer 2019; Bond et al. 2019; Djian et al. 2000; Oviatt 2006; Romero et al. 2016; Schaal 2007; Schmidt 2020; Schmidtler et al. 2015; Tombeil et al. 2020

[2] Adadi & Berrada 2018; Barredo et al. 2020; Bond et al. 2019; Dewey & Wilkens 2019; Ehsan & Riedl 2020; Garcia-Magarino et al. 2019; Gunning et al. 2019; Hayes & Moniz 2020; How 2019; How und Chan 2020; How et al. 2020a; How et al. 2020b; Kahng 2019; Lukowicz 2019; Mucha et al. 2019; Ploug & Holm 2020; Ribera & Lapedriza 2019; Riedl 2019; Springer 2019; Steels 2020; Wang et al. 2019b; Xu 2019; Zanzotto 2019; Zhang et al. 2020; Zhu et al. 2018

[3] Buxbaum & Häusler 2020; Elkmann 2013; Freitag et al. 2016; Häusler & Sträter 2020; Lam 2011; Matthias et al. 2013; Montreuil et al. 2007; Mutlu & Forlizzi 2008; Riener et al. 2005; Riener et al. 2006; Scholtz 2003; Steels 2020

[4] Ahrens 2014; Akata et al. 2020; Bartlett 2006; Beede et al. 2020; Botthoff 2015; Böhmann et al. 2020; Cui & Dai 2008; Dewey & Wilkens 2019; Guszcza et al. 2017; Guszcza 2018; Holzinger et al. 2019; Huchler 2015; Huchler 2016a, 2016b; Huchler & Rhein 2017; Jarrahi 2018; Kaiser & Malanowski 2019; Nahavandi 2019; Nichols et al. 2015; Nisser & Malanowski 2019; Riedl 2019; Roesler & Onnasch 2020; Romero et al. 2016; Schirner et al. 2013; Schmidt 2020; Shneiderman 2020a, 2020b, 2020c; Wang et al. 2019a; Wilkens et al. 2019; Wilkens 2020; Wilson & Daugherty 2018

**Ahrens 2014; Azvine & Wobcke 1998; Bettoni 1995;

[5] Carrico, 2018; Cimini et al., 2020; Hinds et al. 2004; Kaiser und Malanowski 2019; Moniz und Krings 2015; Schaal 2007

Figure 1. *Five perspectives on the meaning of human-centered AI*

The content analysis reveals five basic understandings of how to interpret the human-centricity of AI (see figure 1). Two further perspectives are from AI development. As these perspectives often gain high attention they are integrated on the left hand and right hand side of figure 1 in order to draw an overall picture. This is the understanding of the human being as supplier of data or the definition of the human being as perfect and ideal model for AI development. This means that the human being is central but there is no reflection on a human-centered AI itself and so far a critical reflection whether an AI developed after the model of a human being is human-centered or probably rather to opposite is missing.

The five basic understandings of human-centered AI can be characterized and further exemplified with the help of use cases in the following manner:

(1) There is a **deficit-oriented understanding** of the human being where AI is considered as beneficial and helpful to compensate individual weaknesses and failure in attention, concentration, physical and mental fatigue. AI development aims at making the overall system more robust and failure-adverse. This perspective is a contemporary reinvention of Taylorism with its specific way of thinking and using technology in relation to the individual. A typical use case is assisted driving for train drivers or truck drivers where mental fatigue defines a risk for the driver him- or herself and the environment. The use of AI is for assisting the concentration and drivers behavior through elaborated sensor technology. The core dimension for human-centricity is the compensation of individual weaknesses with the help of AI. Different levels whether system control is always steered by AI or just on individual demand and whether AI can autonomously identify the individual demand and initiate support can be distinguished.

(2) The **data reliability-oriented understanding** gives also emphasis to existing deficits but rather of the AI technology itself instead of the human being. The compensation of existing deficits goes into the direction of making AI better while enhancing its reliability, ease its explainability and foster the individual trust in AI predictions. There is also a focus on fairness as biased data might cause false predictions and estimations. Following this perspective optimization concentrated on AI development but not on changing people. The data reliability-oriented understanding plays an important role e.g. in medical care. As there are many data available from diagnosis there is a potential to use these data for the identification of patterns with the help of machine learning mechanisms (ML) and to improve the accuracy of diagnosis. The correct classification of data plays an important role and requires domain specific knowledge in addition to ML expertise (Dewey & Wilkens 2019). The human-centricity in the data reliability-oriented perspective thus can be indicated by the degree of classifying data with the help of domain experts and thus the interrelatedness between domains.

(3) The **protection-oriented understanding** is widespread in Work Psychology and some parts of Engineering Studies especially when they are related to Work Science or Human Factor / Ergonomics. The physical and mental integrity of the human being is presumed as the highest value in technology development and treated as guideline in job design and technology development. The typical use case is the human computer interaction in production where the intelligent robot assists the human being while performing those tasks which lead to a physical or mental burden (Fahle et al. 2020). The human-centricity in this field depends on the individual autonomy to decide which tasks to be done individually and which automatically and the scope for continuously modifying this decision.

(4) The **potential-oriented understanding** defines an ideal way of combining artificial and individual intelligence. It gives emphasis to a so far unexploited potential of leveraging individual abilities while developing work systems with hybrid intelligence bringing together individual intelligence with AI in a collaborative manner. This is the vision for future job design in Work Science, some fields of Engineering with reference to Industry 4.0 but also in Medicine. There is a strong belief in better outcomes for individual and organizational development as well as task proficiency at the same time. Typical use cases are all fields where decision making needs high accuracy and where the already described challenges of providing reliable data could be solved sufficiently. This is especially decision making in medical diagnosis, therapy or in business development (Dewey & Wilkens 2019; Ellwart et al. 2019). Human-centricity is thus an issue of unfolding individual intelligence.

(5) The **political-oriented understanding** extends the primarily sociological discourse on the distribution of power among different institutions and status groups to the distribution of power between AI and those who use AI in the work context. The normative criteria in use is the clear subordination of technology under the individual control. The normative outline addresses all four understandings introduced before, the deficit-oriented perspective, the data reliability-oriented perspective, the protection-oriented perspective and the potential-oriented perspective as the clarification of the power relationship is considered as critical in all mentioned fields. However, the more political-oriented understanding of human-centricity defines a distinguishable research direction. Use cases for the political-oriented perspective refer to labor regulations and company agreements between employers and work councils which specify the range of control for the work force and the technology. The perspective rather influences the use of technology and not the AI development itself. It is more related to institutional properties defined by industrial relations and can be specified by the level of protection guaranteed for the employees.

Each basic understanding explores its own dimension for evaluating a level of human-centricity of AI in terms of (1) compensating individual weaknesses with the help of AI, (2) enhancing data reliability with the help of domain experts, (3) protecting individual integrity, (4) leveraging individual potential and (5) guaranteeing individual control over AI. The introduced perspectives represent different dimensions of human-centricity of AI and along each dimension there are different maturity levels.

4. Discussion and Outlook

Our literature review revealed different understandings of what human-centricity of AI means. There are at least five dimensions which describe components of human-centricity. The first step of analysis introduced here was important to understand that there are different perspectives (in different disciplines) and to realize that several meanings of human-centricity co-exist. Different developments take place under the same umbrella term. But this first step is not sufficient for drawing a comprehensive picture. There is a need to better understand how these perspectives relate to each other. Further insights are necessary to estimate whether the reliability-oriented perspective and the political-oriented perspective are supposed to define criteria that need to be fulfilled before making use of AI. In this regard reliability would define an AI-inherent criteria, while power distribution would define an AI-external criteria. The three other perspectives might indicate different fields of development which take place in parallel but cannot fully be harmonized as their basic assumptions of the nature of the human being are in contradiction to each other. This is what future research has to explore and discuss in more detail. Moreover, the qualitative hermeneutical approach of literature review presented here should be extended by a more quantitative content analysis which makes use of the key categories identified in this paper. Another important step is to further explore use cases and to describe them with the help of the five dimensions in order to validate these dimensions with respect to their capacity to make clear distinctions between different developments and to indicate different maturity levels.

5. References

Dewey M, Wilkens U (2019) The bionic radiologist: Avoiding blurry pictures and providing greater insights. In: npj Digital Medicine 2(65):1–7. https://doi.org/10.1038/s41746-019-0142-9

Ellwart T, Ulfert A-S, Antoni CH, Becker J, Frings C, Göbel K, Hertel G, Kluge A, Meeßen SM, Niessen C, Nohe C, Riehle DM, Runge Y, Schmid U, Schüffler A, Siebers M, Sonnentag S, Tempel T, Thielsch MT, Wehrt W (2019) Intentional Forgetting in Socio-Digital Work Systems: System Characteristics and User-related Psychological Consequences on Emotion, Cognition, and Behavior. AIS Transactions on Enterprise Systems. 4(1):1-19. https://doi.org/10.30844/aistes.v4i1.16

Fahle S, Prinz C, Kuhlenkötter B (2020) Systematic review on machine learning (ML) methods for manufacturing processes–Identifying artificial intelligence (AI) methods for field application. Procedia CIRP, 93:413-418.

Garcia-Magarino I, Muttukrishnan R, Lloret J (2019) Human-Centric AI for Trustworthy IoT Systems With Explainable Multilayer Perceptrons. In: IEEE Access 7, 125562–125574. https://doi.org/10.1109/ACCESS.2019.2937521

McGregor D (1960) Leadership and Motivation. MIT Press.

Muhle F (2019) Humanoide Roboter als ‚technische Adressen‘ - Zur Rekonstruktion einer Mensch-Roboter-Begegnung im Museum. Sozialer Sinn, 20(1):85–128. https://doi.org/10.1515/sosi-2019-0004

Onnasch L, Maier X, Jürgensohn T (2016) Mensch-Roboter-Interaktion - Eine Taxonomie für alle Anwendungsfälle. Bundesanstalt für Arbeitsschutz und Arbeitsmedizin Dortmund. https://doi.org/10.21934/baua:fokus20160630

Prasad A (2002) The Contest Over Meaning: Hermeneutics as an Interpretive Methodology for Understanding Texts. Organizational Research Methods. 5(1):12-33. https://doi.org/10.1177/1094428102051003

Riasanow T, Setzke DS, Böhm M, Krcmar H (2019) Clarifying the Notion of Digital Transformation: A Transdisciplinary Literature Review. Journal of Competences, Strategy & Management, 10:5-31.

Wilkens U (2020) Artificial intelligence in the workplace - A double-edged sword. In: IJILT 37(5):253–265. https://doi.org/10.1108/IJILT-02-2020-0022

Wilkens U, Lins D, Prinz C, Kuhlenkötter B (2019) Lernen und Kompetenzentwicklung in Arbeitssystemen mit künstlicher Intelligenz. In: Spath, D. & Spanner-Ulmer, B. (Eds.): Digitale Transformation - Gutes Arbeiten und Qualifizierung aktiv gestalten. Gito-Verlag, 71-88.

Wilson HJ, Daugherty PR (2018) Collaborative Intelligence: Humans and AI Are Joining Forces. *Harvard Business Review*, 96(4):114–123.

Xu W (2019) Toward human-centered AI: a perspective from human-computer interactions. In: interactions 26(4):42–46. https://doi.org/10.1145/3328485

The link below gives access to all references cited in figure 1
https://seafile.noc.ruhr-uni-bochum.de/f/d636e976042347b5af09/

Gesellschaft für
Arbeitswissenschaft e.V.

Arbeit HUMAINE gestalten

67. Kongress der
Gesellschaft für Arbeitswissenschaft

Lehrstuhl Wirtschaftspsychologie (WiPs)
Ruhr-Universität Bochum

Institut für Arbeitswissenschaft (IAW)
Ruhr-Universität Bochum

3. - 5. März 2021

GfA-Press

Bericht zum 67. Arbeitswissenschaftlichen Kongress vom 3. - 5. März 2021

Lehrstuhl Wirtschaftspsychologie, Ruhr-Universität Bochum
Institut für Arbeitswissenschaft, Ruhr-Universität Bochum

Herausgegeben von der Gesellschaft für Arbeitswissenschaft e.V.
Dortmund: GfA-Press, 2021
ISBN 978-3-936804-29-4

NE: Gesellschaft für Arbeitswissenschaft: Jahresdokumentation

Als Manuskript zusammengestellt. Diese Jahresdokumentation ist nur in der Geschäftsstelle erhältlich.

Schriftleitung: Matthias Jäger
im Auftrag der Gesellschaft für Arbeitswissenschaft e.V.

Die Verantwortung für die Inhalte der Beiträge tragen alleine die jeweiligen Verfasser; die GfA haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

Screen design und Umsetzung
© 2021 fröse multimedia, Frank Fröse
office@internetkundenservice.de · www.internetkundenservice.de