frontiers | Research Topics

Human-centered AI at work: Common ground in theories and methods

Edited by
Annette Kluge, Corinna Peifer, Uta Wilkens and Verena Nitsch

Coordinated by
Sophie Berretta and Greta Ontrup

Published in
Frontiers in Artificial Intelligence

ISSN 1664-8714
ISBN 978-2-8325-4840-0
DOI 10.3389/978-2-8325-4840-0

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Human-centered AI at work: Common ground in theories and methods

Topic editors
Annette Kluge — Ruhr University Bochum, Germany
Corinna Peifer — University of Lübeck, Germany
Uta Wilkens — Ruhr-University Bochum, Germany
Verena Nitsch — RWTH Aachen University, Germany

Topic coordinators
Sophie Berretta — Ruhr University Bochum, Germany
Greta Ontrup — Ruhr University Bochum, Germany

Citation
Kluge, A., Peifer, C., Wilkens, U., Nitsch, V., Berretta, S., Ontrup, G., eds. (2024). *Human-centered AI at work: Common ground in theories and methods*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-4840-0

Table of contents

frontiers | Frontiers in Artificial Intelligence

TYPE Editorial
PUBLISHED 17 April 2024
DOI 10.3389/frai.2024.1411795

OPEN ACCESS

EDITED AND REVIEWED BY
Dursun Delen,
Oklahoma State University, United States

*CORRESPONDENCE
Annette Kluge
annette.kluge@rub.de

RECEIVED 03 April 2024
ACCEPTED 05 April 2024
PUBLISHED 17 April 2024

CITATION
Kluge A, Wilkens U, Nitsch V and Peifer C (2024) Editorial: Human-centered AI at work: common ground in theories and methods. *Front. Artif. Intell.* 7:1411795. doi: 10.3389/frai.2024.1411795

Editorial: Human-centered AI at work: common ground in theories and methods

Annette Kluge[1]*, Uta Wilkens[2], Verena Nitsch[3] and Corinna Peifer[4]

[1]Organizational and Business Psychology, Ruhr University Bochum, Bochum, Germany, [2]Institute of Work Science, Ruhr University Bochum, Bochum, Germany, [3]Institute of Industrial Engineering and Ergonomics, RWTH Aachen University, Aachen, North Rhine-Westphalia, Germany, [4]Department of Psychology, University of Lübeck, Lübeck, Germany

KEYWORDS

human-AI-teaming, configurational theory, worker participation, work design, social systems, augmentation, team-centered, HRM

Editorial on the Research Topic
Human-centered AI at work: common ground in theories and methods

Human-centered AI at work is being theorized, investigated, and developed in various disciplines providing different definitions and interpretations. Involved disciplines range from information science, machine learning, engineering and robotics, medicine up to ergonomics/work science, psychology, sociology, pedagogics, philosophy, business studies, law and labor relations, just to mention the core disciplines involved in the current debate.

The state-of-the-art presented in the Research Topic's contributions includes lessons learned from socio- technical system design, group work and humane working conditions, negative short-term and long-term consequences in working with automation, design principles of human-autonomy-teaming and effective collaboration between humans collaborating with humans in face of technology, human-machine interaction, workplace democracy and configurational theory.

Authors contribute with reviews, disciplinary and interdisciplinary theory outlines, empirical analysis for tool assessment as well as outcome measures and case illustration. In addition, they provide visionary perspectives to guide future research. The Research Topic includes contributions which (1) systematize the state-of-the-art discourses and methods, (2) specify the operationalization of variables and their relationships, and (3) outline a vision for future practice and related research.

This forms a basis for the development of a research agenda in human-centered AI at work.

State-of-the-art discourses and methods

Berretta et al. conduct a scoping review for a research network analysis and identify five dominant clusters in the field of human-AI-teaming (HAIT) facing (1) human variables, (2) task-dependent variables, (3) AI explainability, (4) AI-driven robotic systems, and (5) effects of AI performance on human perception. It becomes obvious that current research streams are dominated by techno-centric and engineering perspectives but might define a starting point for further elaborating on more human-centric approaches as supported by the authors. They emphasize communication and collaboration requirements in sharing

intents, situational awareness and shared mental models as well as trust among the team members as an issue of HAIT.

Buschmeyer et al. specify the state-of-the-art in the development of methods that are aligned with ISO norms in human-centered design and propose to transfer this framework to AI-based work systems. They introduce a validated instrument assessing (1) system characteristics that are particularly important from the users' perspective; (2) work-related characteristics with respect to mental load and augmentation potential, and (3) cross-task work characteristics. These criteria and the underlying validation define a starting point for future method development.

A research design for measuring the effects of AI tools on human cognitive performance is introduced by Wallinheimo et al.. The authors present a pre-post-measurement among language professionals applying a tool for 5 weeks within a test design. Positive effects for the individual are identified in particular with respect to working memory.

Operationalization of variables and their relationships

Wilkens et al. conduct a cross-disciplinary systematic literature review for specifying criteria as operational benchmarks for human-centered AI at work. In total, they explore eight criteria of human-centricity, (1) trustworthiness and (2) explainability face challenges of technology development, (3) prevention of job loss, (4) health, and (5) human agency & augmentation face challenges of employee development, and (6) compensation of systems' weaknesses, (7) integration of user-domain knowledge, (8) accountability & safety culture reflect challenges of organizational development. With reference to configurational theory the authors argue that different criteria matter in different contexts and depending on stakeholders' responsibility.

Haipeter et al. also contribute toward the contextualization of AI-related research in this field. They refer to the discourse of German speaking sociologists and stress the positive moderator impact of employees' participation in AI implementation as an issue of accountability. The authors illustrate their theoretical argument with a case study description from the German telecommunication industry in which work councils participated in the development of a responsible AI declaration.

Bocklisch and Huchler add further criteria of successful AI implementation for the context of AI-based team settings. Their review among writings from sociology specifies (1) complementarity, (2) shared knowledge & goals, and (3) bounded autonomy as a prerequisite to gain (4) human and team trust in implemented AI.

Mazarakis et al. present a draft for a comprehensive cross-disciplinary model with respect to outcome factors. They pled for the integration of expertise of human factors engineering, human computer interaction, psychology, information science, and adult education in order to envision a future in which AI systems and humans collaborate synergistically to gain higher levels of productivity, innovation, participation and wellbeing.

Vision for future research and practice

Hagemann et al. illustrate hybrid multi-team systems in which human-centered AI emphasize the need for team-centeredness that aligns goals, communication, and decision making with humans. They outline the requirements for such future work contexts with team-centered AI from a sociotechnical perspective, such as cognitive competence, reinforcement learning, and semantic communication.

Fenwick et al. describe the lack of human considerations in HRM tech design and thus develop a vision for the future role of HRM in face of human-AI work systems. They specify the technical, human, and ethical challenges of future HRM systems fully-embedded in a human-centered approach. In this context, they define human-centric AI as AI tools that prioritize and enhance the human experience by making them more intuitive, empathetic, and aligned with human values and needs.

Is there a common ground?

It became clear that a pure focus on technology is too narrow for human-centered approaches but that an exclusive focus on individual variables is also too narrow.

These writings underline that there is a range of criteria indicating human-centered AI at work. The selection of these criteria for empirical analysis varies between disciplines and use fields. It becomes obvious that overall frameworks and criteria exist but need to be adapted to the concrete context as unit of analysis and stakeholders involved whether it is e.g., technology development, human-AI team building or bargaining between status groups.

Hence, it seems that the work system and job characteristics, but especially the team focus and interaction with and around AI which matter as a future unit of analysis. Established methods and leading communities and their impact become clear by the help of these articles.

The Research Topic's contributions show that different perspectives co-exist and—to increase complexity- they co-exist on different levels: individual workplace, team, and organization. On the organizational level Haipeter et al., Wilkens et al., and Fenwick et al. address organizational and social practices of human-centered AI. The team level is addressed in contributions by Hagemann et al., Bocklisch and Huchler, and Berretta et al.. On the workplace level Wallinheimo et al., Mazarakis et al., and Buschmeyer et al. discuss aspects and measurable criteria for designing human-centered workplaces, jobs and AI-assisted tasks.

Hence, based on the Research Topic's contributions, we propose that the common ground of human-centered AI at work

- is embedded in the social systems of an organization, including organizational practices such as HR processes, production processes, and participation processes;
- is value driven- by the striving for decent working conditions (e.g., SDG #8),
- but goes beyond the demand for decent work and sketches images of the augmented worker, working with intelligent

systems that fulfills the requirements of social belonging and relatedness and self-actualization and development;
- acknowledges employees as social beings with needs regarding social contact and motives related to other social beings ("teaming") in the organization;
- augments human capabilities without imposing additional load due to "bad design" in direct human-AI-interaction, and while performing a work task;
- can be assessed and evaluated by means of subjective and objective measures

The Research Topic's visionary contributions underline that human-centered AI needs a focus on interrelated systems to evaluate whether ethical criteria are fulfilled and what are the outcomes and effects on different levels. A set of criteria and variables that needs to be adapted to the use case and unit of analysis were specified. In this way, the research contributions together provide a common ground in human-centered AI at work.

Author contributions

AK: Conceptualization, Writing – original draft, Writing – review & editing. UW: Conceptualization, Writing – original draft, Writing – review & editing. VN: Writing – review & editing. CP: Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. The work on this Research Topic was supported by means of the HumAIne Project and the Project AKzentE4.0.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 25 September 2023
DOI 10.3389/frai.2023.1249322

OPEN ACCESS

EDITED BY
Uta Wilkens,
Ruhr-University Bochum, Germany

REVIEWED BY
Mieczyslaw Lech Owoc,
Wroclaw University of Economics, Poland
Johanna Hartung,
University of Bonn, Germany

*CORRESPONDENCE
Katharina Buschmeyer
katharina.buschmeyer@hs-augsburg.de

RECEIVED 28 June 2023
ACCEPTED 05 September 2023
PUBLISHED 25 September 2023

CITATION
Buschmeyer K, Hatfield S and Zenner J (2023) Psychological assessment of AI-based decision support systems: tool development and expected benefits. *Front. Artif. Intell.* 6:1249322. doi: 10.3389/frai.2023.1249322

Psychological assessment of AI-based decision support systems: tool development and expected benefits

Katharina Buschmeyer[1]*, Sarah Hatfield[1] and Julie Zenner[2]

[1]Faculty of Business, Augsburg Technical University of Applied Science, Augsburg, Germany, [2]Faculty of Liberal Arts and Science, Augsburg Technical University of Applied Science, Augsburg, Germany

This study aimed to develop an evaluation tool that assesses the use of AI-based decision support systems (DSSs) in professional practice from a human-centered perspective. Following the International Organization for Standardization, this perspective aims to ensure that the use of interactive technologies improves users' psychological load experience and behavior, e.g., in the form of reduced stress experience or increased performance. Concomitantly, this perspective attempts to proactively prevent or detect and correct the potential negative effects of these technologies on user load, such as impaired satisfaction and engagement, as early as possible. Based on this perspective, we developed and validated a questionnaire instrument, the Psychological Assessment of AI-based DSSs (PAAI), for the user-centered evaluation of the use of AI-based DSSs in practice. In particular, the instrument considers central design characteristics of AI-based DSSs and the corresponding work situation, which have a significant impact on users' psychological load. The instrument was tested in two independent studies. In Study 1, $N = 223$ individuals were recruited. Based on the results of item and scale analyses and an exploratory factor analysis, the newly developed instrument was refined, and the final version was tested using a confirmatory factor analysis. Findings showed acceptable-to-good fit indices, confirming the factorial validity of the *PAAI*. This was confirmed in a second study, which had $N = 471$ participants. Again, the CFA yielded acceptable-to-good fit indices. The validity was further confirmed using convergent and criterion validity analyses.

KEYWORDS

AI-based decision support systems, work, human-centered evaluation, survey inventory, system properties, characteristics of the supported task, psychological load

1. Introduction

Professionals have to make various decisions during the course of their work. For example, asset managers must choose between various investment options, whereas lawyers have to decide on a possible defense strategy for a particular case. For a decision to be made, a conscious and voluntary choice must be made among several alternative courses of action by comparing, considering, and evaluating them based on available data, information, and knowledge (Büssing et al., 2004; Rau et al., 2021). Owing to the growing amount of data and information in our increasingly digitalized and globalized world, decision-making processes have become very complex across various professions (Latos et al., 2017; van Laar et al., 2017; Timiliotis et al., 2022). For many, keeping a track of all relevant new data and information

when making decisions and placing them in the context of existing knowledge poses a great challenge (Timiliotis et al., 2022). The amount of available data in some areas has become so vast that it cannot be processed by humans, as it simply exceeds their information processing capacity (Koltay, 2017; Saxena and Lamest, 2018; Shrivastav and Kongar, 2021). Moreover, in everyday work, highly complex decision-making situations are often complicated by stressors, like an elevated time and performance pressure. Such challenges and, for many professionals, overstraining decision-making situations lead to higher levels of uncertainty, stress, and lower decision quality (Phillips-Wren and Adya, 2020). They also affect, for example, job satisfaction (Nisar and Rasheed, 2020) and organizational productivity (Miller and Lee, 2001; Vosloban, 2012).

Although the intensification of digitization and globalization leads to increased risks for companies and professionals, it also opens up new opportunities. For example, the accumulation of data, both in terms of quantity and quality, has enabled impressive improvements in the field of artificial intelligence (AI), which helps in the development of extremely powerful algorithms (Nicodeme, 2020). They are often based on machine-learning models, which are more scalable and flexible than traditional statistical models (Rajula et al., 2020), making them appropriate tools for today's dynamic and complex work environments. For problems such as those described above, researchers have acknowledged the particular great potential of the use of AI-based decision support systems (DSSs; see Brynjolfsson et al., 2011; Cai et al., 2019; Shin, 2020; Tutun et al., 2023), which are often referred to as *Augmented Intelligence Systems* (Jarrahi, 2018; Hassani et al., 2020; Walch, 2020; Kim et al., 2022). As the name suggests, these systems are designed to augment, and not replace, humans in complex decision-making situations by taking over specific task components, like processing big data, which are difficult for human intelligence to handle. In professional practice, this looks like this: AI-based applications analyze the huge amounts of data and information available and make hidden patterns in the data accessible to humans in the form of insights or concrete recommendations for action (Konys and Nowak-Brzezińska, 2023). Humans are free to decide whether to follow the system's recommendation. Thus, humans remain the central element in the interpretation and verification of AI-based systems, resulting in complex decision-making situations and continued sovereignty over the final decisions and associated actions (Hellebrandt et al., 2021). This is pivotal because even though there are powerful algorithms behind AI-based DSSs, they also have limitations and weaknesses like overfitting, lack of transparency, and biases (Pedreschi et al., 2019). Humans can compensate for these weaknesses through their inherent strengths and mental acumen (e.g., critical thinking, creativity, and intuition; Spector and Ma, 2019; Wilkens, 2020). Hence, the introduction of augmented intelligence systems ideally leads to a synergetic interaction between human and machine intelligence, which helps professionals to better handle increased cognitive demands (Kirste, 2019). Consequently, they feel appropriately challenged and less burdened in work-related decision-making situations (Cai et al., 2019), which is reflected, for example, in their higher task performance (Li et al., 2021). From a business perspective, the improved decision-making process should, for example, lead to increased company's performance (Brynjolfsson et al., 2011). To summarize, the introduction of an AI-based DSS should create a mutually beneficial scenario for professionals and their companies.

However, it has been noted that many AI initiatives have failed to achieve their objectives. This can be attributed to several reasons, including technical challenges like insufficient databases, organizational failures like inadequate expectation management, and failed system design. For example, users often cannot find a new system that is sufficiently useful or transparent, making them unwilling to use the system (Westenberger et al., 2022). To avoid this, since 2019, the International Organization for Standardization (ISO) has advocated the adoption of a human-centered design approach in the development of interactive systems such as AI-based DSSs. The approach "aims to make systems usable and useful by focusing on users, their needs and requirements, and applying knowledge and techniques from the fields of human factors/ergonomics and usability. This approach increases effectiveness and efficiency; improves human wellbeing, user satisfaction, accessibility, and sustainability; and counteracts the potential negative effects of use on human health, safety, and performance" (ISO International Organization for Standardization, 2019). To achieve this, the ISO International Organization for Standardization (2019) recommends an organizations' active user involvement throughout the development process and follows a four-step design process: (1) understanding and describing the context of use, (2) specifying user requirements, (3) developing design solutions, and (4) evaluating the design solutions. The fourth step of evaluation plays a decisive role in this process. Here, the success of the project is determined; if not successful, stakeholders can study the concrete modification measures required and process steps that must be repeated. However, when assessing success, the ISO International Organization for Standardization (2019) also underpins the importance of observing not only whether system introduction has led to the intended effects but also whether possible negative, unintended side-effects have occurred.

Based on previous project reports, it is evident that it is common for the introduction of AI-based systems to lead to negative, unintended side-effects. For example, in a case study in the banking sector, Mayer et al. (2020) observed that the introduction of an AI-based system in the lending department led to a perceived loss of competence and reputation among system users. To derive appropriate actions in cases where unintended outcomes occur and in those where desired outcomes are not achieved, it is necessary to gain an accurate understanding of the impact of a new system (and its individual characteristics) on the relevant work situation and its users. This consideration is particularly important when an AI-based DSS is assisting with a core activity, and showcases that the introduction of AI-based DSSs carries particular weight in influencing the user's load experience—both in desirable and undesirable ways. A well-known example of this is related to service and customer support professionals, whose core activity is dealing with customer issues on a daily basis. These professionals now increasingly have access to AI-based DSSs that assist them with a relatively high degree of automation: it provides them with concrete suggestions for actions regarding the requests made by customers. In this scenario, the need for a thorough and comprehensive evaluation of system implementation is undeniable. However, practical evaluation instruments considering this holistic

perspective are currently lacking. The available assessment methods comprise either (a) user experience surveys, which enable the evaluation of the impact of specific system properties on users (e.g., SUS, Bangor et al., 2008; Perceived Usefulness and Ease of Use scales, Davis, 1989; meCue, Minge et al., 2017) or (b) job analyses, which allow a comprehensive examination of the influence of new technologies as a whole on task design (e.g., WDQ, Morgeson and Humphrey, 2006; TBS-GA(L), Rudolph et al., 2017; FGBU, Dettmers and Krause, 2020). However, to the best of our knowledge, there is currently no specific practical assessment tool for evaluating the use of AI-based DSSs and that effectively combines both levels of consideration (i.e., user experience and job analyses). Therefore, this study aims to close this gap by developing and validating a questionnaire instrument that not only captures the properties of an AI-based DSS but also the characteristics of the corresponding work situation, thus providing a holistic evaluation framework.

2. Conceptualization and use of the evaluation instrument

The newly developed evaluation questionnaire, called Psychological Assessment of AI-based DSSs (*PAAI*), is based on the core idea of many occupational psychology models (e.g., the job demand control model, Karasek, 1979; the Stress-Strain model, Rohmert., 1984; the Transactional Model of Stress and Coping, Lazarus and Folkman, 1984; and the Action Regulation theory, Hacker, 1978). These models describe that when assessing work tasks, a distinction should be made between *work characteristics* (trigger factors) and the resulting *psychological load* (trigger reactions) experienced by professionals. Work characteristics include all identifiable aspects of a task, such as complexity, social environment, and work equipment, which affect human engagement in the task. The immediate impact of psychological work characteristics on an individual, considering one's internal (e.g., intelligence) and external resources (e.g., social support), is referred to as psychological load. Depending on the alignment between psychological work characteristics and individual resources, the psychological load can be positive (e.g., activation and flow experience) or negative (e.g., mental overload and stress). Persistent psychological load has medium- and long-term consequences, including positive outcomes, like satisfaction and wellbeing, or negative outcomes, like dissatisfaction and reduced performance (ISO International Organization for Standardization, 2019).

From the above-mentioned perspective, the introduction of a new work tool (e.g., an AI-based DSS) can be perceived as a new work task characteristic that should help to alleviate user psychological overload. Following Hacker (1978) hierarchical levels of technology-based tasks, this impact on psychological load can occur at three levels (see Figure 1).

First, the immediate interaction with the system can trigger a psychological response in the user. Depending on the design of the system and the user's resources (e.g., technical knowledge), this can be positive (e.g., enjoyment), or negative (e.g., irritation). Previous research on conventional information systems and DSSs show which design features are particularly influential for user experience (see Venkatesh and Davis, 2000; Calisir and Calisir, 2004; Alshurideh et al., 2020). These findings also apply to AI-based DSSs (see Henkel et al., 2022; Meske and Bunde, 2022). However, since they differ from conventional DSSs in their probabilistic nature and black-box character (Zhang et al., 2020; Jussupow et al., 2021), related design features must also be considered in this case. Section 2.1 provides an overview of the traditional as well as the specific characteristics that current research suggests promote positive human-augmented intelligence interactions.

Second, the implementation of AI-based DSSs can influence users' psychological load regarding the entire processing of the task supported by the system. To this end, they are implemented to support their users in task processing and to handle other—in this case unfavorable—work characteristics (e.g., information overload). Thus, DSSs can and should reduce, for example, user uncertainty and stress in task processing. DSSs based on AI methods are currently seen as particularly promising tools for this purpose. This is because their extremely powerful algorithms enable them to capture both highly complex and dynamic problems (Kirste, 2019; Kim et al., 2022). However, it is likely that the great power of the systems will not only have a direct impact—by providing immediate support in decision-making situations—but also an indirect impact on the psychological load of users. Because they also have a great potential—targeted or untargeted—to change the character of the supported work tasks from the professionals' point of view, which in turn determines their load experience. This kind of change is desirable when, for example, it enables professionals to perceive previously overwhelming tasks as less complex and therefore less stressful. On the other hand, unintended consequences can occur if, for example, employees perceive fewer opportunities for further learning and, as a result, the personality-enhancing aspect of the work activity is lost.

Third, to make the levels of consideration complete, the introduction of an AI-based DSS should also be considered from the overall job perspective. This is because the introduction of an AI-based DSS can also potentially affect cross-task work characteristics, consequently affecting the psychological load of users in relation to their jobs as a whole. For example, recent research shows that professionals can feel threatened in their jobs by the introduction of new technologies (Gimpel et al., 2020); for example, this may occur if new technologies cause them to perceive their jobs as less future-proof (Lingmont and Alexiou, 2020).

To sum up, the introduction of an AI-based DSS can influence the psychological load of users (and the associated consequences) at three different levels: (1) during immediate human-AI interaction, (2) during processing of the supported task, and (3) during executing the entire job. Since the three levels are interrelated, they should all be considered in an evaluation of these systems. During the evaluation, it may turn out that the users' psychological load or related consequences on one or more levels deviate from the desired result, or it may be grasped that further optimization potential remains to be identified. In these cases, it is advisable to first take a closer look at the lowest level, examining its characteristics and derive possible required modification needs. Thereafter, the other levels can be gradually included in the analysis. In this way, the

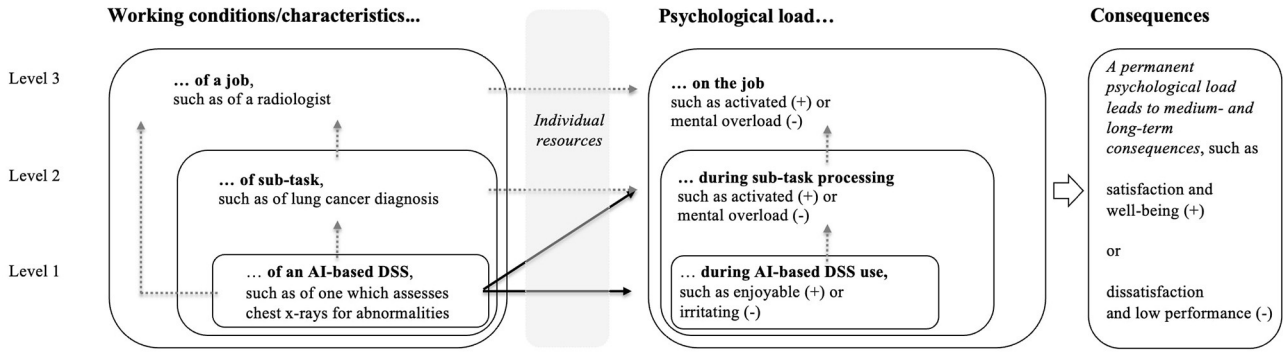

FIGURE 1
The basic model of the assessment instrument psychological assessment of AI-based decision support systems (PAAI).

need for action can be identified from a more specific to a more general level.

Moreover, if the various work characteristics are surveyed both as part of a one-time measurement after its introduction and before system introduction, the evaluation can also examine how these have changed as a result of the initiative. Thus, all effects of system introduction—including unintended side-effects—can be precisely tracked and easily corrected if necessary. In practice, this means that the system evaluation can be applied both as part of a one-time measurement or in the form of a pre-post measurement, depending on the exact purpose of the evaluation. In the next section, we provide discussions on level-specific characteristics that (a) have a significant impact on the psychological load of professionals and (b) are closely associated with the implementation of AI-based DSSs. They form the assessment measures in the *PAAI*.

2.1. Level 1: human-AI-interaction

At the finest level of consideration, the focus is on the individual design characteristics of AI-based DSSs that strongly influence users' psychological load during their interactions with the system and their willingness to use it in the first place. The best-studied and most important characteristics that all types of information systems should satisfy—and thus the characteristics that are most consistently evaluated—are *Perceived Usefulness* and *Perceived Ease of Use*, as described in the Technology Acceptance Model (Davis, 1989). When evaluating these two factors for AI-based DSSs, it is also important to consider the unique properties of these systems, which significantly influence user perceptions of *Perceived Usefulness* and *Perceived Ease of Use*.

More specifically, the *Perceived Usefulness* of a system is strongly influenced by its information quality (Machdar, 2019) and thus, in the case AI-based DSSs, by the accuracy of the system. This is because, as mentioned earlier, AI-based DSSs operate on a probabilistic basis, meaning that there is no absolute guarantee that a system result will be correct (Zhang et al., 2020). Thus, in order for users to experience systems as valuable and trustworthy, they must provide correct results with the highest possible probability (Shin, 2020). Therefore, when evaluating the *Perceived Usefulness* of AI-based DSSs, it is important to ask directly about the perceived quality of the system output, in addition to considering other conventional and concrete design features, such as the task-technology fit of the system (Goodhue and Thompson, 1995).

The *Perceived Ease of Use* of an AI-based DSSs, as with conventional DSSs, is generated by design features such as self-descriptiveness (ISO International Organization for Standardization, 2020) and simplicity (Lee et al., 2007). AI-based DSSs should also pay particular attention to ensuring that the system output is presented in a way that is easy to understand (Henkel et al., 2022). The aforementioned probabilistic nature of AI systems can make it difficult for users to correctly interpret system output; research in cognitive psychology has shown that humans often have difficulty correctly understanding probabilities, which lead to increased misjudgments (Anderson, 1998). Therefore, to facilitate good user interactions, it is necessary to create intuitively designed interfaces and present results in a human-centered way to reduce the risk of misinterpretation.

In addition to the peculiarity of the probabilistic nature and associated uncertainty of AI-based DSSs, they differ from conventional systems that they develop their own programming rules. Their algorithmic mechanisms for model generation are therefore not transparent (Jussupow et al., 2021). As a result, the underlying logic of these systems is often referred to as a black-box model (Kraus et al., 2021). The lack of information about why an AI-based system operates in a certain way also complicates the interpretation of system outputs. Therefore, an increasing number of AI-based DSSs provide additional explanations using Explainable AI (XAI) methods. These relate to how an AI-based system arrives at its output and what goes into that output (Arrieta et al., 2020). However, a failure to sufficiently *perceive these explanations as comprehensible* can negatively impact trust and acceptance of AI-based DSSs (Shin et al., 2020), as well as the cognitive effort required for decision-making (Meske and Bunde, 2022).

Freely accessible AI assistance systems, like ChatGPT (https://chat.openai.com) and DeepL (https://www.deepl.com/translator), are increasingly bringing AI to the forefront of public awareness.

Simultaneously, our own interactions with these cloud-based solutions highlight the criticality of their reliable accessibility, particularly in hectic working situations. Frequent unavailability (e.g., due to overwhelming user demand) can lead to stress (Körner et al., 2019). Therefore, *Perceived Availability* is also a central influencing factor of user experience.

Table 1 provides an overview of what existing research has revealed about the influences of the four system characteristics—*Perceived Usefulness*, *Perceived Ease of Use*, *Perceived Comprehensibility*, and *Perceived Availability*—on users' psychological loads. Below, a detailed description of each construct considered in the inventory is presented.

Perceived Usefulness refers to the extent to which an individual believes that using an AI-based DSS improves decision-making effectiveness and efficiency (Davis, 1989; Krieger and Lausberg, 2021). To perceive a system as useful, a high task suitability must be perceived by the users, which implies that the system meets the specific requirements of the task it supports (Goodhue and Thompson, 1995; Klopping and McKinney, 2004). Furthermore, the system must provide high-quality information and deliver accurate, timely, complete, and relevant results (Gorla et al., 2010; Hsiao et al., 2013; Atta, 2017; Machdar, 2019).

Perceived Ease of Use encompasses the extent of effortlessness of use of an AI-based DSS as perceived by individuals (Davis, 1989). To achieve this, systems should be designed with clear functions and user-friendly interfaces to ensure that the system output is easy to understand (Doshi-Velez and Kim, 2017; Iriani and Andjarwati, 2020; Sati and Ramaditya, 2020). Additionally, simplicity should be prioritized in system design, which can be achieved through the reduction, organization, integration, and prioritization of system features (Lee et al., 2007).

Perceived Comprehensibility includes individuals' perceptions of the extent of their clear understanding of reasons for the output generated by the system (Coussement and Benoit, 2021). To achieve this, it is advantageous to provide both general model explanations, which elucidate functional relationships between the input and output variables, and specific explanations, which aid in understanding individual data-related outputs (Kraus et al., 2021). The users should not be overwhelmed with excessive system details; instead concise and effective information should be offered that enables them to effectively utilize the system within their task environment (Mercado et al., 2016).

Perceived Availability encompasses the extent to which individuals perceive the content of a system as reliably accessible and retrievable. Usually, this is affected by factors such as the frequency of unexpected system updates, system crashes, error messages, and technical problems (Körner et al., 2019).

2.2. Level 2: AI-supported task

At the second evaluation level, along with the system characteristics discussed at Level 1, specific task characteristics of the supported tasks were considered. To facilitate the evaluation process, three broad groups of task characteristics were identified: requirements, resources, and stressors (Iwanowa, 2006). The task characteristics considered in the inventory for each of these groups are discussed in detail below, and they have the following common traits: (a) they have a significant impact on the psychological load of professionals and (b) it is very likely that the introduction of AI-based DSSs will affect them directly or the way professionals interact with them (see Table 2).

2.2.1. Requirements

The group of *requirements* includes all work characteristics that professionals must meet in order to successfully and effectively perform their work tasks. Therefore, requirements are inherent to the nature of the task and unavoidable. In general, the characteristics of this group are considered positive and beneficial for personality development *per se*, but only as long as they fit individual resources of the jobholder. Otherwise, it leads to psychological underload or overload (Iwanowa, 2006; Semmer and Zapf, 2018). Two requirements were considered in the developed inventory: *Perceived Complexity and Decision-making Requirements and Perceived Cooperation and Communication Requirements*.

Perceived Complexity and Decision-making Requirements refer to the perceived level of mental demand of a task. It can be categorized into various levels, ranging from routine activities with rehearsed mental requirements to activities requiring productive thinking and problem-solving (Hacker, 2016). Decision-making is a component of complex tasks, and its degree can be assessed by various measures like the number of variables involved (Stemmann and Lang, 2014).

Perceived Cooperation and Communication Requirements involve the perceived need to inform and coordinate with colleagues. It includes factors like the duration of communication, number of partners involved, mode of communication (direct or indirect), and content, like information sharing, instruction dissemination, and collaborative problem-solving (Richter et al., 2014). These requirements are often accompanied by highly complex tasks, as they often necessitate cooperation and collaboration between different specialists or departments owing to the diverse skills and knowledge required (Helquist et al., 2011).

2.2.2. Resources

The group of *resources* includes all work characteristics that provide opportunities for action and may or may not be used voluntarily (Zapf, 1998; Semmer and Zapf, 2018). However, the professionals should be aware of these possibilities. Resources have a predominantly positive relationship with the indicators of maintaining and promoting health and fostering personal development (Iwanowa, 2006). In the evaluation tool, the two resources of *Perceived Latitude for Activity and Perceived Use of Qualifications and Learning Opportunities* should be considered and explained.

Perceived Latitude for Activity is a multidimensional construct that includes the perceived scope of action, design, and decision-making in a professional set-up. Scope of action refers to the range of available action-related options, including the choice of approach, resources, and temporal organization of task components. This defines the degree of flexibility in performing subtasks in a professional scenario. Design latitude refers to the

TABLE 1 PAAI's Level 1 assessment criteria and their influence on professionals' psychological load.

Evaluation criteria by system level	Associated results on ...	
	... professionals' experience of human-machine-interaction (Level 1)	... professionals' psychological load during task processing (Level 2)
Perceived usefulness (PU)	Attitude toward use (Alhashmi et al., 2020); behavioral intention to use a system (Venkatesh and Davis, 2000; Alhashmi et al., 2020; Al Shamsi et al., 2022); technology trust (Amin et al., 2014); technology satisfaction (Amin et al., 2014); actual usage (Rigopoulos et al., 2008)	Decision quality (Wook Seo et al., 2013); performance (Omar et al., 2019; Arnold et al., 2020); engagement (Lackey et al., 2016); workload (Lackey et al., 2016)
Perceived ease of use (PEU)	Attitude toward use (Alhashmi et al., 2020); behavioral intention to use a system (Venkatesh and Davis, 2000; Alhashmi et al., 2020; Al Shamsi et al., 2022); technology trust (Amin et al., 2014); technology satisfaction (Amin et al., 2014); actual usage (Rigopoulos et al., 2008)	Performance (Omar et al., 2019); mental effort (Lackey et al., 2016); frustration (Lackey et al., 2016)
Perceived comprehensibility (PC)	Technology trust (Shin, 2020, 2021; Liu et al., 2022); perceived value (Liu et al., 2022); perceived quality of advice (Gaube et al., 2023)	Performance (Stowers et al., 2020; Gaube et al., 2023); workload (Mercado et al., 2016); cognitive effort (Meske and Bunde, 2022)
Perceived availability (PA)	Stress (Körner et al., 2019); perceived value (Baldauf et al., 2020; Prakash and Das, 2020)	Stress (Körner et al., 2019)

ability to design processes independently based on goals. Decision-making latitude considers the degree of decision-making authority in task definition and delineation and determines the degree of autonomy associated with an activity (Ulich, 2011).

Perceived Use of Qualifications and Learning Opportunities refers to the perception that one can optimally utilize own existing expertise, skills, and abilities professionally. Therefore, a process of learning maintenance occurs. Conversely, low levels of this resource indicated unlearning (Büssing et al., 2004). Learning opportunities are closely related to the use of own qualifications and resources in executing job responsibilities. Interestingly, the existence of learning opportunities can only be assessed by comparing existing and required knowledge, skills, and abilities (Rau et al., 2021).

2.2.3. Stressors

The group of *stressors* encompasses all factors that impede the achievement of task goals and those that require professionals to make additional efforts or take additional risks. These efforts and risks, in turn, increase their work load, time, and effort (Büssing et al., 2004; Semmer and Zapf, 2018). Thus, dealing with stressors has adverse effects on the mental health of most professionals (Iwanowa, 2006). In the *PAAI*, four stressors are considered: *Perceived Information Overload, Perceived Lack of Information, Perceived Time and Performance Pressure, and Perceived Qualification Deficits*.

Perceived Information Overload involves the perception of the need to consider or evaluate an amount of information that is larger than the one's information intake and processing capacity (Dettmers and Krause, 2020). Furthermore, it has been noted that humans have a unique characteristic: the more information we are offered, the more information we think we need (Krcmar, 2011). According to Heinisch (2002), this leads to a paradox in knowledge society. This states that in the midst of the flood of information, there is a lack of information.

Perceived Lack of Information indicates that information is perceived as missing, unavailable, or not up to date (Dettmers and Krause, 2020).

Perceived Time and Performance Pressure describes the perceived imbalance between three work components, as follows: quantity, time, and quality (Trägner, 2006). The mismatch between these three components lies in the fact that a certain amount of work cannot be accomplished in the required or necessary quality in the available working time (Schulz-Dadaczynski, 2017).

Perceived Qualification Deficits indicate that, from the perspective of professionals, the work task assigned to them does not match their existing qualifications; these qualifications include technical competencies (e.g., specialized knowledge, work techniques, skills, and abilities) and social and communicative competencies required for the proper execution of a task (Richter et al., 2014). However, the mismatch can be attributed to qualification deficits for an activity, for example, due to insufficient training; as a result, workers feel overtaxed. Low qualification adequacy can also be seen when workers perform activities below their qualification level, triggering a qualitative underchallenge (Dettmers and Krause, 2020).

2.3. Level 3: overall job

The third level considers the workplace's cross-task characteristics, which extend beyond task-related aspects and affect users' psychological load in relation to their job. As mentioned earlier, in AI implementation projects, there is a risk that the introduction of AI-based systems may induce a higher level of *Perceived Job Insecurity* among its users. This fear is related to concerns about job loss owing to automation or insufficient proficiency in using digital technologies and media (Gimpel et al., 2020). According to a recent study by Lingmont and Alexiou (2020), professionals who are highly aware of AI and robotics tend to perceive a higher level of job insecurity than those with lower awareness. The implementation of AI-based

TABLE 2 PAAI's Level 2 assessment criteria and their influence on professionals' psychological load, and the potential influence of AI-based DSSs on these criteria.

Characteristics group	Evaluation criteria on task level	Associated outcomes on professionals' psychological load and load consequences	Possible positive intended effect through system implementation	Possible negative unintended effect through system implementation
Requirements	Perceived Complexity and decision-making requirements (PCDR)	Stress (Phillips-Wren and Adya, 2020); Performance (Maynard and Hakel, 1997; Mosaly et al., 2018; Chinelato et al., 2019); mental effort (Mosaly et al., 2018); satisfaction (Morgeson and Humphrey, 2006)	With AI-based system support, professionals probably perceive tasks/decisions as less complex or feel more confident in dealing with them (Mayer et al., 2020; Wanner, 2021). As a result, they feel less stressed (Cai et al., 2019; Lee et al., 2021). The improved data-driven decision basis is also expected to improve decision quality and performance (Li et al., 2021; Wanner, 2021).	There is the risk of automation bias, in that users may come to often rely on the system's advice and not critically reflect on it (Skitka et al., 1999; Mayer et al., 2020; Panigutti et al., 2022). They may then perceive the task as not complex enough or monotonous, which can lead them to feel under challenged and bored (Loukidou et al., 2009).
	Perceived cooperation and communication requirements (PCCR)	Stress (Zeffane and McLoughlin, 2006); mental health (Lu and Argyle, 1991); happiness (Lu and Argyle, 1991)	If employees feel more confident in decision-making situations with system support, the need for cooperation with colleagues is likely to decrease. If these were previously perceived as too high and high losses of time and energy were associated with them, this can have a positive effect on the experience of psychological load.	If employees feel more confident in decision-making situations with system support, the need for exchange with colleagues is likely to decrease. If these were previously perceived as appropriate, this can be perceived as negative, since social exchange reduced, for example.
Resource	Perceived Latitude for activity (PLA)	Work engagement (Dettmers and Krause, 2020); satisfaction (Morgeson and Humphrey, 2006); motivation (Glaser et al., 2015); loss of irritation (Glaser et al., 2015; Dettmers and Krause, 2020); loss of psychosomatic complaints (Dettmers and Krause, 2020)	No effects are expected.	The introduction of new technologies like AI is often accompanied by process standardization (Silva and Gonçalves, 2022), which in turn probably limit professionals' perceived latitude for activity.
	Perceived use of qualifications and learning opportunities (PUQL)	Satisfaction (Rowden and Conine, 2005); engagement (Jin and McDonald, 2017); intention to stay (Steil et al., 2020)	No effects are expected.	There is a risk that by using AI-based DSSs, professionals rely little on their own skills and thus lose their expertise over time. Since there is no maintenance learning and they do not take advantage of learning opportunities (Mayer et al., 2020).
Stressors	Perceived information overload (PIO)	Irritation (Dettmers and Krause, 2020); stress (Misra et al., 2020); tension (Theron, 2014); tiredness (Theron, 2014); loss of job satisfaction; (Theron, 2014); decision quality (Hwang and Lin, 1999)	Through the use of AI-based DSSs that bundle and process information, the information overload should perceived to be less or its handling easier due to the new resource (Maes, 1995; Aussu, 2023).	Often, stressors influence psychological load to such an extent that they can overshadow other work characteristics (Phillips-Wren and Adya, 2020). Therefore, there is an AI-based DSSs will have little or no impact on professionals' psychological load experience if they remain too high.
	Perceived lack of information (PLI)	Irritation (Dettmers and Krause, 2020); psychosomatic complaints (Dettmers and Krause, 2020)	By using the system, it is likely for fewer information deficits to occur, as the system generates new patterns, new information, and insights from the data (Haefner et al., 2021).	

(Continued)

10.3389/frai.2023.1249322

TABLE 2 (Continued)

Characteristics group	Evaluation criteria on task level	Associated outcomes on professionals' psychological load and load consequences	Possible positive intended effect through system implementation	Possible negative unintended effect through system implementation
	Perceived time and performance pressure (PTPP)	Irritation (Dettmers and Krause, 2020); psychosomatic complaints (Dettmers and Krause, 2020); exhaustion (Syrek et al., 2013); loss of work–life balance (Syrek et al., 2013)	It can be assumed that the use of AI-based DSSs alleviates the time and performance pressure experienced by professionals. This is primarily attributed to the system enabling employees to make faster and more confident decisions. Alternatively, a similar level of pressure may be perceived, but the associated challenges can be better managed by professionals through the use of the aforementioned technology (Wanner, 2021; Tutun et al., 2023).	
	Perceived qualification deficits (PQD)	Irritation (Dettmers and Krause, 2020); psychosomatic complaints (Dettmers and Krause, 2020); loss of work engagement (Dettmers and Krause, 2020)	It is likely that AI-based systems compensate for existing skill deficiencies of professionals (Gaube et al., 2023) consequently reducing the experience of negative load consequences.	

DSSs probably increases users' awareness of AI, which in turn could raise concerns related to job insecurity and thereby increase psychological load.

3. Scale and item generation and qualitative review

To measure the 13 characteristics described (see Section 2), we developed the respective *PAAI*. To be able to observe the impact of the constructs on the professionals in the context of an evaluation, it is also necessary to collect appropriate indicators of the professionals' psychological load and related consequences. Fortunately, several scales already exist for this purpose (e.g., NASA-TLX, Hart and Staveland, 1988; irritation scale, Mohr et al., 2006; stress experience, Richter, 2000), from which a particular variable can be selected as per project objectives. However, one exception is the measurement of psychological load during immediate human-AI interaction. To the best of our knowledge, no questionnaires are available for this variable as of yet. Therefore, we developed an additional scale to measure user *Irritation during System Use*. In developing the 14 scales, we took care to keep the number of statements per scale as short as possible while ensuring that a minimum of three items met the scientific validity criteria (Mvududu and Sink, 2013). The items were formulated using generic terms, so that they can be applied to different occupations and types of AI-based DSSs. Simultaneously, the assessment incorporates design recommendations to facilitate the identification of specific causes and derivation of appropriate action measures.

To test the clarity and face validity of the developed items, cognitive pretests were conducted with $N = 10$ individuals without prior experience of AI-based DSSs in an occupational context. First, a paraphrasing method was used (Porst, 2013). In this method, participants were asked to reproduce the individual statements of the questionnaire in their own words. If a respondent did not understand a statement or understood it incorrectly, the statement was rephrased, clarified with examples, or removed. The revised items were then tested in the second step with the remaining respondents. In this step, a sorting technique was used to examine how respondents assigned the given items to the given constructs. It could be said that the sorting technique is a type of factor analysis that does not require previously collected data (Porst, 2013). All items assigned to the correct category by at least 75% of raters were retained. This development process resulted in a questionnaire with 59 items (Table 1; Supplementary material).

4. Empirical testing of the developed items and scales

The newly developed items and scales were empirically tested in two consecutive studies using causal samples. Adjustments like the deletion of items were made as required. Study 1 began with an analysis of the items and scales. Subsequently, an exploratory factor analysis (EFA) was conducted to examine the factor structure of the questionnaire derived from the item and scale analysis. The resulting factor structure, which was expected to be statistically

and theoretically adequate, was subjected to confirmatory factor analysis (CFA) for further validation, employing the maximum likelihood (ML) estimation method, and allowing the factors to correlate (detailed information is accessible in the provided data on OSF—see the data availability statement). Study 2 tested the factorial validity of the final model from Study 1 using CFA, utilizing maximum likelihood estimation, with a different sample size. Furthermore, the final scale was assessed in terms of its convergent and predictive validity. In both cases, correlations with other variables collected simultaneously using established instruments were analyzed. Correlation analyses, item and scale analyses, and EFA were conducted using IBM SPSS software version 26.0. For the CFA, RStudio software (version 4.2.0) was used.

4.1. Study 1

4.1.1. Method

4.1.1.1. Participants and procedures

On January 24, 2023, $N = 250$ participants from the UK were recruited from a crowdsourcing platform, Prolific, for the survey. The prerequisite was that the participants must being employed and regularly use a DSS in the job. For the survey, it did not matter whether the DSS is based on AI methods because the properties surveyed can be assessed for all DSSs, regardless of which technical solutions are behind them. After excluding participants who, for example, missed one of the two attention checks or had superficial response patterns, $N = 223$ participants remained ($n = 132$ female, $n = 91$ male). Most respondents (42.6%) were aged between 30 and 39 years and worked in customer service and support (12.1%), organization, data processing and administration (11.7%), and marketing and sales (11.2%). Regarding educational level, the most had a bachelor's degree (46.6%).

4.1.1.2. Materials

The survey comprised the newly developed instrument (see Supplementary Table 1) and general demographic questions (e.g., age, gender, and field of activity). Aside from the general demographic questions, all other items were answered on a 5-point scale ranging from 1 (doesn't apply at all) to 5 (applies completely).

4.1.2. Analysis and results

4.1.2.1. Item and scale analysis

In the context of item and scale analysis, it is first checked whether the items are too easy, too difficult, or insufficiently differentiated, and then the internal consistency of the scales is assessed. Specifically, items with a difficulty index between $P = 0.20$–0.80 and a discriminatory power of $rit \geq 0.40$ are targeted (Bortz and Döring, 2016; Kalkbrenner, 2021). All but two items (19 and 52) met these criteria, and the two items that did not meet the criteria were therefore deleted. Supplementary Table 2 presents the descriptive statistics for each scale at the end of the item and scale analyses.

4.1.2.2. EFA

In the next step, the scales or items were further tested separately within their corresponding levels, as described herein: (1) human-AI interaction, (2) AI-supported task, and (3) overall job (as described in Section 2). We conducted an EFA for each. The prerequisites for the EFA needed to be examined further. First, inter-item correlations were checked; that is, whether each item correlated with at least three other items with a value between $r = 0.20$ to $r = 0.85$. Moreover, we tested whether each item had a Kaiser-Meyer-Olkin (KMO) measure of $\geq$0.70, Bartlett's test for sphericity of $p < 0.05$, and a measure of sampling adequacy (MSA) >0.8 (Kalkbrenner, 2021; Zhang et al., 2021). These conditions were fulfilled in all cases. To determine the most suitable number of factors for summarizing the elements within each level, two analyses were conducted: principal component analysis (PCA) and parallel analysis (Kalkbrenner, 2021). The results in Supplementary Table 3 indicate that different criteria suggest different numbers of factors.

In subsequent EFAs, the modeling process employed principal axis factorization (PFA) with the oblique rotation method Promax because of the assumed inter-factor correlations (Moosbrugger and Kelava, 2008). The initial starting point for each analysis was the highest assumed number of factors, as shown in Supplementary Table 3, to ensure maximum information preservation. Throughout the exploration, individual items were systematically excluded, and after each exclusion, the PFA was recalculated. Items were excluded if at least one of the following three conditions was not met: First, items should exhibit a factor loading of at least $\lambda \geq 0.40$. Second, it is desirable that the items have no cross-loadings, that is, they should have no loadings of $\lambda \pm 0.40$ on two or more factors. Meanwhile, if an item does demonstrate cross-loadings, but the loading on one factor is $\lambda \geq 0.10$ higher than on the other factors, researchers can use their theoretical understanding to decide whether the variable should be assigned to a factor or deleted. Third, from a theoretical perspective, all items should fit the respective factors to which they were assigned according to the EFA (Korner and Brown, 1990; Kalkbrenner, 2021). Consequently, of the 28 items included in the EFA at Level 1 (human-AI interaction), seven items (items 1, 5, 9, 12, 13, 17, and 22) were deleted. At Level 2 (AI-supported task), of the 26 items included in the EFA, eight items (items 32, 36, 39, 40, 47, 48, 49, and 50) were deleted. According to the criteria mentioned above, Item 30 should also have been deleted as it loaded highly on two factors (difference $\lambda = 0.05$). However, this was not done because of (a) strong theoretical reasons and (b) the decision is statistically confirmed in the subsequent CFA because the model quality is better with the inclusion of the item in the scale than without it.

For the remaining 21 items at Level 1, a five-factor structure resolved a total of $R^2 = 60.29\%$ of the variance. In terms of content, the five factors were consistent with the theory described in Section 2. However, at Level 2, contrary to the theoretical assumption, a six-factor structure—not an eight-factor structure—was determined for the remaining 18 items, and it explained $R^2 = 60.12\%$ of the cumulative variance. Specifically, the EFA results indicated that the items developed to measure the constructs of decision-making and complexity requirements should be combined with those representing the construct of information overload into a single factor. This merger was supported by theoretical considerations. Moreover, all items related to the construct "lack of information" were deleted because of their high cross-loadings. This deletion was

TABLE 3 Overview of exploratory factor analysis and confirmatory factor analysis results in Study 1.

Level 1: Human-AI-interaction				Level 2: AI-supported task				Level 3: Overall job			
Factor	Item	Load EFA	Load CFA	Factor	Item	Load EFA	Load CFA	Factor	Item	Load EFA	Load CFA
PU	2	0.74	0.75	PCDR	30	0.70	0.80	PJI	57	0.89	0.89
	3	0.82	0.79		31	0.72	0.71		58	0.89	0.89
	4	0.77	0.78		45	0.69	0.62		59	0.70	0.70
	6	0.78	0.77		46	0.71	0.63				
	7	0.75	0.77	PCCR	33	0.80	0.83				
PEU	8	0.76	0.76		34	0.77	0.76				
	10	0.77	0.74		35	0.88	0.84				
	11	0.70	0.67	PLA	37	0.57	0.59				
	14	0.74	0.77		38	0.78	0.67				
	15	0.74	0.75		41	0.72	0.77				
PC	16	0.83	0.76	PUQL	42	0.73	0.75				
	18	0.71	0.70		43	0.73	0.76				
	20	0.69	0.70		44	0.86	0.81				
	21	0.69	0.74	PTPP	51	0.75	0.75				
PA	23	0.59	0.62		53	0.76	0.76				
	24	0.78	0.76	PQD	54	0.87	0.87				
	25	0.78	0.78		55	0.83	0.83				
	26	0.77	0.76		56	0.77	0.77				
ISU	27	−0.89	0.89								
	28	−0.91	0.91								
	29	−0.82	0.82								
Model fit indices:				**Model fit indices:**				**Model fit indices:**			
$\chi^2/df = 1.48$				$\chi^2/df = 1.48$				$\chi^2/df = 0$			
CFI = 0.97				CFI = 0.97				CFI = 1.00			
TLI = 0.96				TLI = 0.96				TLI = 1.00			
RMSEA = 0.05				RMSEA = 0.05				RMSEA = 0.00			
SRMR = 0.04				SRMR = 0.05				SRMR = 0.00			

N = 223.

χ^2, Chi squared; df, degree of freedom; CFI, Comparative Fit Index; ISU, Irritation during System use; PA, Perceived Availability; PC, Perceived Comprehensibility; PCCR, Perceived Cooperation and Communication; PCDR, Perceived Complexity and Decision-making Requirements; PEU, Perceived Ease of Use; PJI, Perceived Job Insecurity; PLA, Perceived Latitude for activity; PQD, Perceived Qualification deficit; PTPP, Perceived Time and performance pressure; PU, Perceived Usefulness; PUQL, Perceived Use of Qualifications and Learning Opportunities; RMSEA, Root Mean Square Error of Approximation; SRMR, Standardized Root Mean Residual; TLI, Trucker–Lewis index.

10.3389/frai.2023.1249322

theoretically justifiable and appropriate. For the three remaining items at Level 3, as theoretically hypothesized, we obtained a one-factor structure that resolved $R^2 = 68.95\%$ of the variance. A comprehensive depiction of the factor structure encompassing the items and their corresponding factor loadings for each scale is provided in Table 3.

4.1.2.3. CFA

CFAs were conducted to confirm the factor structure of the developed instrument according to the EFAs. This analysis was used to determine the fit between the model and obtained data (Bandalos, 2002). Following Hu and Bentler (1999), the model fit index was determined using chi-square/degree of freedom (good fit $= 0 \leq \chi^2/df \leq 0.2$; acceptable fit $= 2 < \chi^2/df \leq 0.3$), comparative fit index (CFI; good fit $= 0.97 \leq$ CFI ≤ 1.00; acceptable fit $= 0.95 \leq$ CFI < 0.97), Tucker-Lewis index (TLI; good fit $= 0.97 \leq$ TLI ≤ 1.00; acceptable fit $= 0.95 \leq$ TLI < 0.97), root mean square error of approximation (RMSEA; good fit $= 0 \leq$ RMSEA ≤ 0.05; acceptable fit $= 0.05 <$ RMSEA ≤ 0.08), and standardized RMR (SRMR; good fit $= 0 \leq$ SRMR ≤ 0.05; acceptable fit $= 0.05 <$ SRMR ≤ 0.10). The indices for the three models are listed in Table 3. Acceptable to good fit indices were observed for all models.

4.2. Study 2

4.2.1. Method

4.2.1.1. Participants and procedure

From 6 to 16 February 2023, a second survey was conducted in the UK and US via Prolific. The participation criteria were the same as those of Study 1. Of the $N = 535$ participants, $N = 471$ individuals ($n = 235$ female, $n = 233$ male, $n = 3$ non-binary) remained after data cleaning as in the first study. Most respondents (36.5%) were between 30 and 39 years and had a bachelor's degree (39.7%). In addition, most worked in customer service and support (12.1%), organization, data processing and administration (8.1%), and marketing and sales (7.9%).

4.2.1.2. Materials

Along with the final questionnaire inventory from Study 1 (Supplementary Table 4) and demographic data, we collected data on load indicators and consequences to test criterion validity. Specifically, system use satisfaction was surveyed through a custom-developed item, "I like using the system" and trust in a system was surveyed using a custom-developed question, "How much do you trust the system?" The subjective stress experienced during a task was surveyed using the six items developed by Richter (2000), which were translated from German to English. Moreover, mental effort and mental exhaustion were assessed with two questions from the BMS short scales, which were translated from German into English (Debitz et al., 2016). Task enjoyment was surveyed with the item "How much pleasure do you usually get from the work task?" Furthermore, competence experience during task processing was assessed using four items adapted from a prior study (Sailer, 2016). Job satisfaction was surveyed using an item, which was translated from German into English, from a past study (Kauffeld and Schermuly, 2011). For later testing of convergent validity, the six-item meCue (Minge et al., 2017), which measures the usefulness and usability of technologies, was used. All statements were answered on a 5-point scale ranging from 1 (doesn't apply at all) to 5 (applies completely). The questions, for example on mental effort and exhaustion, were responded on a slider scale ranging from 1 (low) to 10 (high).

4.2.2. Analysis and results

4.2.2.1. Item and scale analysis

Table 4 shows the descriptive statistics for all the scales. All scales showed an internal consistency of $\alpha \geq 0.70$ (Hussy et al., 2013), and the *Perceived Time and Performance Pressure* scale a Spearman-Brown coefficient of 0.757. In addition, high mean values and low standard deviations were observed for most scales.

4.2.2.2. CFA

To test the factor structure of the final version of the *PAAI* on another sample, a second CFA was conducted for each Level and for the overall instrument. As in Study 1, the results showed acceptable to good fit indices (Hu and Bentler, 1999) for all three models per Level, and for the overall model (see Table 5). A detailed overview of the individual factor loadings is provided in Supplementary Table 4.

4.2.2.3. Correlation analysis

A correlation analysis was conducted to further test the validity of the *PAAI* (see Supplementary Tables 5, 6). In this analysis, the Level 1 scales were correlated with the meCue (Minge et al., 2017) to test their convergent validity and overall summary. It showed that all Level 1 scales as a whole correlate strongly with the meCue ($r = 0.83$, $p < 0.01$) according to Cohen's (1988) criteria ($r \geq 0.10$ small, $r \geq 0.30$ moderate, $r \geq 0.50$ strong effect). From the perspective of individual scales, the *Perceived Usefulness* ($r = 0.70$, $p < 0.01$) and *Perceived Ease of Use* ($r = 0.83$, $p < 0.01$) showed particularly very high correlations, as expected, and the *Perceived Comprehensibility* ($r = 0.54$, $p < 0.01$) and *Perceived Availability* ($r = 0.56$, $p < 0.01$) showed high correlations.

Moreover, for all levels, the collected criterion-related variables capturing the load indicators and the corresponding predictive scales were found to be correlated. These results demonstrate the criterion-related validity of the instrument. Furthermore, in the task characteristic group of requirements, the *Perceived Complexity and Decision-making Requirements* scale was moderately positively related to mental effort ($r = 0.41$, $p < 0.01$) and mental exhaustion ($r = 0.28$, $p < 0.01$), and simultaneously positively related to task enjoyment ($r = 0.15$, $p < 0.01$). Contrastingly, stressors such as *Perceived Qualification Deficits* correlated with negative load indicators like stress experience ($r = 0.64$, $p < 0.01$), and had no significant relationship with positive load indicators, such as task enjoyment ($r = 0.06$, not significant) or competence experience during task processing ($r = -0.30$, $p < 0.01$). This undesirable influence of stressors was also evident at Level 3, with the *Perceived Job Insecurity* scale correlating negatively with job satisfaction ($r = -0.34$, $p < 0.01$). However, variables in the resources group such as *Perceived Latitude for Activity* (Level 2) were moderately correlated with task enjoyment ($r = 0.40$, $p < 0.01$) and competence experience during task processing ($r = 0.31$, $p < 0.01$).

TABLE 4 Results of the item and scale analysis in Study 2.

Level	Characteristics or load	Scale	Number of items	Range of the scale	Cronbach's α	Mean value	Standard deviation
1	System characteristics	*PU*	5	1;5	0.84	3.85	0.64
		PEU	5	1;5	0.84	3.67	0.69
		PC	4	1;5	0.77	3.38	0.73
		PA	4	1;5	0.83	3.63	0.77
		meCue	6	1;5	0.86	3.72	0.67
	Load indicators	*ISU*	3	1;5	0.88	2.22	0.85
	Task characteristics	Satisfaction with system use	1	1;5	-	3.50	0.92
		Trust in the system	1	1;10	-	7.15	1.60
2		*PCDR*	4	1;5	0.72	3.20	0.78
		PCCR	3	1;5	0.87	2.96	1.01
		PLA	3	1;5	0.76	2.94	0.90
		PUQL	3	1;5	0.82	3.21	0.90
		PTPP	2	1;5	0.76	2.95	0.99
		PQD	3	1;5	0.76	2.19	0.87
	Load indicators	Stress experience	6	1;5	0.86	2.09	0.74
		Mental effort	1	1;10	-	6.76	1.89
		Mental exhaustion	1	1;10	-	5.35	2.17
		Task enjoyment	1	1;10	-	5.54	2.25
		Competence experience in task processing	4	1;5	0.77	3.66	0.69
3	Job characteristics	*PJI*	3	1;5	0.88	2.23	1.02
	Load indicators	Job satisfaction	1	1;10	-	6.92	2.08

N = 471.

ISU, Irritation during System use; PA, Perceived Availability; PC, Perceived Comprehensibility; PCCR, Perceived Cooperation and Communication; PCDR, Perceived Complexity and Decision-making Requirements; PEU, Perceived Ease of Use; PJI, Perceived Job Insecurity; PLA, Perceived Latitude for activity; PQD, Perceived Qualification deficit; PTPP, Perceived Time and performance pressure; PU, Perceived Usefulness; PUQL, Perceived Use of Qualifications and Learning Opportunities.

PAAI criteria are shown in italics.

TABLE 5 Confirmatory factor analysis results from Study 2.

	χ^2/df	CFI	TLI	RMSEA	SRMR
Good fit	$0 \leq \chi^2/df \leq 0.2$	$0.97 \leq \text{CFI} \leq 1.00$	$0.97 \leq \text{TLI} \leq 1.00$	$0 \leq \text{RMSEA} \leq 0.05$	$0 \leq \text{SRMR} \leq 0.05$
Acceptable fit	$0.2 < \chi^2/df \leq 0.3$	$0.95 \leq \text{CFI} < 0.97$	$0.95 \leq \text{TLI} < 0.97$	$0.05 < \text{RMSEA} \leq 0.08$	$0.05 < \text{SRMR} \leq 0.10$
Models tested					
Level 1 scales	1.76	0.97	0.97	0.04	0.04
Level 2 scales	2.04	0.96	0.95	0.05	0.04
Level 3 scales	0	1.00	1.00	0.00	0.00
All scales	1.45	0.96	0.96	0.03	0.04

N = 471.

χ^2, Chi squared; df, degree of freedom; CFI, Comparative Fit Index; RMSEA, Root Mean Square Error of Approximation; SRMR, Standardized Root Mean Residual; TLI, Trucker–Lewis index.

5. Discussion

This study aimed to develop and validate an evaluation tool that assesses the use of AI-based DSSs in the workplace, strongly emphasizing the human aspect. Using this human-centered perspective, this study ultimately aimed to ensure that the implementation of new technology has a positive impact on user psychological wellbeing, as well as helps in avoiding unintended negative consequences that could hinder personal development in the workplace. To be able to verify the outcomes of AI-based DSSs implementation, it was necessary to understand the effects of system deployment on users and the associated work situation in

a differentiated manner as part of the evaluation. Only in this way can the need for adaptation be specifically derived if necessary.

Thus, an instrument called *PAAI* was developed to capture the following design characteristics in the context of AI-based DSSs: (1) system characteristics of AI-based DSSs that are particularly important from the users' perspective; (2) work-related characteristics of the AI-supported task that are particularly influential for professionals' psychological load and known to play a frequent role in the context of implementation of new technology based on the augmented intelligence approach; and (3) cross-task work characteristics that are often relevant from the professionals' perspective in this context. The selection of the specific system-, task-, and job-related design characteristics collected in the *PAAI* is guided by this research. In total, 13 characteristics were initially identified from the literature and measured with 56 items after an initial, concise, cognitive preliminary study. The newly developed questionnaire was then extensively tested in a preliminary quantitative study, which yielded the necessary adjustments to the instrument. The refined version of the questionnaire was tested in a second quantitative study using another sample. The final instrument encompasses a total of 11 design characteristics measured by 39 items (see Supplementary Table 4).

Looking at the results

In both studies, the items or scales generated to capture system-, task-, or job-related characteristics were first analyzed in detail within their associated characteristic group or level of consideration. This procedure guaranteed that the newly developed scales function well within their respective levels, and before they are evaluated as a coherent whole across all three levels at the end of Study 2. It also ensured that the three questionnaire parts can be used independently if needed.

In the first section of the questionnaire, in which items assessed the perceived system characteristics of AI-based DSSs from the user's perspective, the EFA in Study 1 yielded a four-factor structure: *Perceived Usefulness, Perceived Ease of Use, Perceived Comprehensibility, and Perceived Availability*. It fit well with the expected theoretical structure (see Section 2). Strictly speaking, however, the EFA resulted in a five-factor structure, as items measuring the psychological load indicator of professionals (i.e., in the form of *Irritation during System Use*) were also included in the analysis. This is because no questionnaire had, thus far, been developed to capture users' direct experiences of psychological load during direct human-AI interactions. Therefore, the inclusion of the *Irritation during System Use* scale in the EFA was necessary to ensure the validity of this newly developed instrument. Furthermore, an exploratory analysis showed that for the items on system characteristics, the EFA results do not differ depending on whether the *Irritation during System Use* scale is included, which indicates the high separability of this dependent variable from the other four independent variables. Contrarily, this construct of independence within the scales of system properties was not as pronounced as shown by the cross-loadings observed in the EFA for the associated items. This finding is not unexpected, as previous studies have shown a close association between individual system characteristics, such as perceived usefulness and ease of use (Suki and Suki, 2011). To delineate the individual constructs more clearly, we eliminated items that could not be clearly assigned to a particular construct. This item reduction procedure was not problematic, as the analysis commenced with an item surplus in order to identify the most powerful and relevant items for the primary study. Therefore, despite the item reduction, all Level-1-scales (a total of five) show good internal consistency, as indicated by the Cronbach's alpha values (Hussy et al., 2013) ranging from $\alpha = 0.81–0.91$ in Study 1. These acceptable reliabilities are also shown in Study 2, as the Cronbach's alpha values ranged from $\alpha = 0.77–0.88$.

Importantly, both studies had high mean values (tending to be at the high end of the scale) and low standard deviations for all four system characteristic scales. These observations can be interpreted as an indication of the strong predictive role of the captured system characteristics for system use. This is because participants in both studies evaluated only systems that are regularly used in everyday life. Furthermore, the criterion validity of the developed system characteristic scales on professionals' psychological load experience during immediate human-AI interactions was also empirically confirmed in this study. In particular, in Study 2, all four scales showed, following Cohen's (1988) methodology, moderately negative correlations with the criterion *Irritation during System Use*; thus, both the criterion and construct validity of the newly developed scales were demonstrated. As in Study 1, consistently acceptable goodness-of-fit indices were observed for the final scales in the CFA on a second independent sample. Along with factorial validity, convergent validity was also examined to confirm construct validity. As expected, the *Perceived Usefulness* and *Perceived Ease of Use* scales showed a strong correlation with the *meCue* (Minge et al., 2017), as this well-established instrument maps two very similar constructs. Conforming to this, the *Perceived Comprehensibility* and *Perceived Availability* scales correlated moderately with the *meCue* scale, as with the *Perceived Usefulness* and *Perceived Ease of Use* scales.

The second part was designed to obtain the relevant work-related characteristics of AI-supported tasks in the context of AI-based DSSs. Initially, this questionnaire section comprised eight task characteristics based on the theoretical foundations (see Supplementary Table 1), and had a total of 27 items. However, the assumed factor structure for the newly developed items could not be confirmed in Study 1. The EFA results revealed high cross-loadings between items in the hypothesized scales of *Perceived Complexity and Decision-Making Requirements, Perceived Information Overload, Perceived Lack of Information, and Perceived Time and Performance Pressure*, suggesting the need for adjustment. These findings can be attributed to the fact that work-related characteristics often co-occur in practice and partially influence each other (Dettmers and Krause, 2020; Phillips-Wren and Adya, 2020; Rau et al., 2021).

To ensure the valid measurement of the constructs, two approaches were employed. First, inaccurate items were gradually eliminated during the exploratory phase. This was not problematic, because the preliminary study started with a larger number of items than required. Moreover, whenever reasonable, items from

closely related constructs were combined into a single factor. Consequently, during the item deletion process, all items related to the construct of *Perceived Lack of Information* were removed, along with single items from other assumed scales. By removing the entire *Perceived Lack of Information* construct, the relationships among the remaining constructs became clearer. Given that AI-based DSSs are primarily designed to improve the handling of information overload and complex decision processes (Dietzmann and Duan, 2022; Stenzl et al., 2022), and addressing information lack through the identification of novel patterns only yields the possibility of an incidental benefit, we do not consider the omission of this task characteristic of the inventory to be critical. The second approach for improvement involved merging two closely related constructs: *Perceived Complexity and Decision-Making Requirements* and *Perceived Information Overload*. This decision to include the *Perceived Information Overload* scale in the *Perceived Complexity and Decision-making Requirements* is also supported by theoretical considerations since high decision-making and complex requirements often involve dealing with a substantial number of variables and their associated information (Phillips-Wren and Adya, 2020; Rau et al., 2021). From a professional perspective, these two components are likely to be perceived as unified entities rather than as separate constructs. After these adjustments, the CFA results of both the preliminary and main studies consistently showed acceptable to good fit indices for the new six-factor structure of the questionnaire. Furthermore, satisfactory reliabilities were observed for all scales in Studies 1 and 2, ranging from $\alpha = 0.72$ to $\alpha = 0.87$. It is noteworthy that although the *Perceived Time and Performance Pressure* scale comprises only two items, it met all the reliability and validity criteria. Since the survey was to be as concise as possible for practical reasons, there was no need to add a third item to the scale, as often recommended by prior studies like that conducted by Mvududu and Sink (2013). In addition to factorial validity, the criterion validity of all the scales was confirmed in the main study. For example, the mental effort criterion showed the strongest correlation with the scale *Perceived Complexity and Decision-making Requirements*. This result is consistent with previous research, showing that complexity and decision demands are significant predictors of cognitive effort (Lyell et al., 2018). Furthermore, the correlation results support those of previous studies, demonstrating a significant relationship between users' negative load experiences and qualification deficits (Dettmers and Krause, 2020). As expected, the results also showed the positive correlation of the two resources variables of *Perceived Latitude for Activity* and *Perceived Use of Qualifications and Learning Opportunities* with indicators of positive load, like task enjoyment.

To holistically evaluate the introduction of an AI-based DSS, the last section of the *PAAI* focuses on a cross-task job characteristic, more specifically, on *Perceived Job Insecurity*. This focus serves to enable the tool to provide data on whether this variable increases from the perspective of the affected professionals as a result of the introduction of new technology. In both the preliminary and main studies, the results consistently confirmed the reliability and construct validity of the scale. Furthermore, the main study proves criterion validity, as this construct correlates negatively with positive load indicators, in this case job satisfaction, in line with previous reports.

Thus, the results of Studies 1 and 2 provide compelling evidence of the validity and reliability of all three parts of the final questionnaire. Furthermore, the construct validity of the questionnaire instrument was assessed in the main study using CFA, showing satisfactory-to-very good fit indices and the overall validity of the instrument.

5.1. Limitations and future research

Thus far, the empirical results have confirmed the reliability and validity of the new questionnaire instrument, which can be used both in practice in the context of evaluating AI-based DSSs and in scientific research. This instrument is economic, easy-to-use, and has a solid scientific basis. However, this study has some limitations. First, the validation of the questionnaire relied solely on questionnaire-based instruments; therefore, the results may have been influenced by social desirability bias. Therefore, future validation studies should include a wider range of data sources. For example, the additional assessment of relevant system properties using mathematical metrics, such as system accuracy, is an important sub-design criterion for perceived usefulness (Yin and Qiu, 2021) psychological load indicators using physiological and biochemical measures (Lean and Shan, 2012). Second, because we aimed at developing a questionnaire with the shortest possible survey duration, it is critical to note that convergent validity was only tested for the characteristic scales of the lowest human-AI-interaction level. Scholars are thus recommended to test both the convergent and discriminant validity of the scales at the other two levels (AI-supported tasks and overall jobs) in future studies. This would allow for a more comprehensive assessment of instrument validity across all system levels. Third, data collection for this study was conducted via a paid crowdsourcing platform, which may raise concerns about data quality (Douglas et al., 2023).

To address these issues, (a) a conscious decision was made to use a sample provider that, according to prior research (Peer et al., 2022), delivers the highest data quality; (b) two control questions were included in the questionnaire used in each of the studies, and the data of all participants who failed one of these questions were immediately excluded. Nonetheless, it is advisable to validate the instrument using a separate sample that does not receive financial compensation for participation as well as that is not exclusively from the US and the UK—but instead from various countries and cultures. This will further strengthen the confidence in the observed results and their generalizability beyond the paid crowdsourcing platform sample. Furthermore, the target group of the questionnaire comprised people who used DSSs in their daily work. We decided to not impose any further specific participation conditions related to the AI methods behind the system for several reasons. First, the *PAAI* can be used separately from this specific technical solution, even if it is simultaneously assumed that, especially in the case of AI-based DSSs, attention must be paid to ensure that systems are designed to be, e.g., sufficiently comprehensible because of

their black-box nature. Second, the definition of the subject of AI varies widely, and until this date, there is no universally expected definition of the subject (Alter, 2023). Finally, users may not be aware of the specific technical solutions underlying their DSS. Therefore, by omitting the technical solution, the *PAAI* could focus on capturing user perceptions and experiences of the DSS, rather than their awareness of the underlying AI methods. Nevertheless, future studies should investigate whether the term "AI" alone influences user experience and behavior. Previous research investigating the impact of AI-based DSSs on user psychological load has so far mainly focused on the particular characteristics of accuracy and transparency of these systems (see Stowers et al., 2020; Jacobs et al., 2021; Jussupow et al., 2021; Gaube et al., 2023). These arise, as described in the theory section, from the essence of AI-based DSSs; namely, their probabilistic nature and black-box character. In addition, individual studies have been exploring the effect of the timing of the introduction of AI-generated advices to users (Jussupow et al., 2021; Langer et al., 2021). For example, initial findings suggest that users are more satisfied and experience higher self-efficacy in task processing if they receive support from the system after first independently processing the information underlying their decision-making (Langer et al., 2021). However, further evidence is required before generalizations can be made on this topic. Future research efforts should therefore explore other design and implementation characteristics of AI-based DSSs, including the timing of support, and investigate the influence of the term "AI". This work can yield deeper understandings of the dynamics of user experience and psychological load in relation to AI-based DSSs. Ultimately, this research may enable the identification of other potential key characteristics that should be considered when evaluating AI-based DSSs. Furthermore, future research could conduct research to identify under which design aspects and contextual conditions users of augmented intelligence systems (e.g., AI-based DSSs) perceive these systems as optimal complements to their own abilities, and how the degree of augmentation affects users' psychological load; for example, regarding their own experience of motivation and empowerment. Finally, it would also be interesting to investigate what inversely influences professionals' experience of load on their perception of the system and work-related characteristics. Previous studies have indicated that professionals' current load experience also influences their perceptions of working conditions (Rusli et al., 2008). To increase the acceptance of users toward a new work system, it could be helpful to implement it not in particularly stressful peak periods but in times of moderate workloads.

5.2. Practical implications

In augmented-intelligence projects, it is critical for organizations to prioritize future users throughout the development and validation processes. As mentioned in the introduction, the success of augmented intelligence implementation, like AI-based DSSs, in the workplace ultimately depends on the experience and behavior of the employees involved. If they are not ready to use the new technology, the project is likely to fail during the implementation phase. To avoid this, organizations could follow the four phases of the human-centered design approach when implementing an AI-based DSS (ISO International Organization for Standardization, 2019). This requires the involvement of a transdisciplinary team that includes psychological experts and usual technical experts. The expertise of the former is valuable in tasks like requirements analysis, adaptation of work habits, changes in communication during implementation, and evaluation. In the evaluation, it is not sufficient to assess only whether the intended effects were achieved; it is equally important to identify any unintended effects that may have occurred. To gain insight into why newly implemented systems actually have an impact, organizations could conduct comprehensive surveys to understand their effects on individuals and their work environments. This comprehensive understanding facilitates the development of tailored action plans. Along with the use of the *PAAI* evaluation tool, organizations could consider incorporating complementary criteria of system performance, like accuracy (Kohl, 2012) and response time (Tsakonas and Papatheodorou, 2008). Furthermore, Organizations could also include users' individual resources in the evaluation, considering the stress and strain models (ISO International Organization for Standardization, 2017). For example, their expertise level or AI knowledge (Gaube et al., 2023) could be included to identify users' qualification needs. Moreover, data on project management related variables could be collected. As per prior studies, changes in communication are extremely influential in AI project success, particularly expectation management (Alshurideh et al., 2020). Therefore, organizations could also evaluate the success of specific communication measures and whether any further action is needed.

6. Conclusion

This study emphasizes the importance of a human-centered design approach in the development and implementation of augmented intelligence projects, as well as the implementation of a user-centered evaluation within this framework. An evaluation tool suitable for this purpose, called *PAAI*, was developed. The novel instrument can be seen as a holistic tool that, along with immediate interface design, focuses on personality-promoting workplace design. The *PAAI* can not only be used selectively to evaluate the impact of AI-based DSSs implementation projects on users, but also as a starting point for the requirements analysis of the four-step human-centered design process. Thus, it could be used as a pre-post measurement. Thus, organizations can use the *PAAI* to develop AI-supported workplaces that are conducive to a positive mental health among workers. Although the *PAAI* was validated by the two independent studies reported in this manuscript, further research is required to collect data from more diverse samples and verify evidence consistency. Moreover, the use of additional data sources, such as objective

and qualitative measures, should help further validate the newly developed instrument.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://osf.io/vfykx/?view_only=03628f3f25c249668e0770f0fb6f9c6f.

Ethics statement

Ethical approval was not required for the studies involving humans because the study was a voluntary survey on Decision Support Systems and Working Conditions that did not give rise to an ethics vote (e.g., there was no deception). The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements because after the information text about the survey, there was a button to agree to the survey. Thus, the consent was only in the form of a click, but without a signature. This was because the survey was completely anonymous and without any form of intervention. Anyone could voluntarily participate in the study and stop at any time.

Author contributions

KB was responsible for the conception and design of the study, data collection, analysis, interpretation, and was the primary writer of the manuscript. SH accompanied with the conception of the study. SH and JZ assisted with the study design and critically revised the manuscript. All authors approved the final version of the manuscript for submission.

Funding

The research presented in this study has been carried out within the Research Project AIXPERIMENTATIONLAB (Project No. EXP.01.00016.20).

Acknowledgments

The authors gratefully acknowledge the support of the German Federal Ministry of Labor and Social Affairs (BMAS) and the Initiative New Quality of Work (INQA).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2023.1249322/full#supplementary-material

References

Al Shamsi, J. H., Al-Emran, M., and Shaalan, K. (2022). Understanding key drivers affecting students' use of artificial intelligence-based voice assistants. *Educ. Inf. Technol.* 27, 8071–8091. doi: 10.1007/s10639-022-10947-3

Alhashmi, S. F. S., Salloum, S. A., and Abdallah, S. (2020). "Critical success factors for implementing artificial intelligence (AI) projects in Dubai Government United Arab Emirates (UAE) health sector: Applying the extended technology acceptance model (TAM)," in *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2019*, eds. A. E. Hassanien, K. Shaalan, and M. F. Tolba (Cham), 393–405.

Alshurideh, M., Al Kurdi, B., and Salloum, S. A. (2020). "Examining the main mobile learning system drivers' effects: a Mix empirical examination of both the expectation-confirmation model (ECM) and the technology acceptance model (TAM)," in *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2019*, eds. A. E. Hassanien, K. Shaalan, and M. F. Tolba (Cham), 406–417.

Alter, S. (2023). "How can you verify that I am using AI? Complementary frameworks for describing and evaluating AI-based digital agents in their usage contexts," in *56th Hawaii Conference on System Sciences (HICSS)*.

Amin, M., Rezaei, S., and Abolghasemi, M. (2014). User satisfaction with mobile websites: the impact of perceived usefulness (PU), perceived ease of use (PEOU) and trust. *Nankai Bus. Rev. Int.* 5, 258–274. doi: 10.1108/NBRI-01-2014-0005

Anderson, J. L. (1998). Embracing uncertainty: the interface of Bayesian statistics and cognitive psychology. *Conserv. Ecol.* 2, 102. doi: 10.5751/ES-00043-020102

Arnold, A., Dupont, G., Kobus, C., Lancelot, F., and Liu, Y.-H. (2020). "Perceived usefulness of conversational agents predicts search performance in aerospace domain," in *Proceedings of the 2nd Conference on Conversational User Interfaces* (New York, NY: Association for Computing Machinery), 1–3.

Arrieta, B. A., Díaz-Rodríguez, N., Ser, D. J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inform. Fusion* 58, 82–115. doi: 10.1016/j.inffus.2019.12.012

Atta, M. T. (2017). The effect of usability and information quality on decision support information system (DSS). *Arts Soc. Sci. J.* 8, 257. doi: 10.4172/2151-6200.1000257

Aussu, P. (2023). "Information overload: coping mechanisms and tools impact," in *Research Challenges in Information Science: Information Science and the Connected World. Proceedings: 17th International Conference, RCIS 2023 Corfu, Greece, May 23–26, 2023*, eds. S. Nurcan, A. L. Opdahl, H. Mouratidis, and A. Tsohou (Cham: Springer), 661–669.

Baldauf, M., Fröehlich, P., and Endl, R. (2020). "Trust me, I'm a doctor – User perceptions of AI-driven apps for mobile health diagnosis," in *Proceedings of the 19th International Conference on Mobile and Ubiquitous Multimedia*, eds. J. Cauchard, M. Löchtefeld, MUM '20 (New York, NY: Association for Computing Machinery), 167–178.

Bandalos, D. L. (2002). The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling. *Struct. Equ Model. Multidiscip. J.* 9, 78–102. doi: 10.1207/S15328007SEM0901_5

Bangor, A., Kortum, P. T., and Miller, J. T. (2008). An empirical evaluation of the system usability scale. *Int. J. Hum. Comput. Interact.* 24, 574–594. doi: 10.1080/10447310802205776

Bortz, J., and Döring, N. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften. 5th Edn.* Berlin and Heidelberg: Springer.

Brynjolfsson, E., Hitt, L. M., and Kim, H. H. (2011). *Strength in Numbers: How Does Data-driven Decisionmaking Affect Firm Performance?* Berlin. Available online at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1819486 (accessed June 28, 2023).

Büssing, A., Glaser, J., and Höge, T. (2004). Psychische und physische Belastungen in der ambulanten Pflege: Ein Screening zum Arbeits- und Gesundheitsschutz. *Z. Arbeits Organisationspsychol.* 48, 165–180. doi: 10.1026/0932-4089.48.4.165

Cai, C. J., Reif, E., Hegde, N., Hipp, J., Kim, B., Smilkov, D., et al. (2019). "Human-centered tools for coping with imperfect algorithms during medical decision-making," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, eds. S. Brewster, G. Fitzpatrick, A. Cox, V. Kostakos, CHI '19 (New York, NY: Association for Computing Machinery), 1–14.

Calisir, F., and Calisir, F. (2004). The relation of interface usability characteristics, perceived usefulness, and perceived ease of use to end-user satisfaction with enterprise resource planning (ERP) systems. *Comput. Hum. Behav.* 20, 505–515. doi: 10.1016/j.chb.2003.10.004

Chinelato, R. S. d. C., Ferreira, M. C., and Valentini, F. (2019). Work engagement: a study of daily changes. *Ciencias Psicol.* 13, 3–8. doi: 10.22235/cp.v13i1.1805

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* New York, NY: Routledge.

Coussement, K., and Benoit, D. F. (2021). Interpretable data science for decision making. *Decis. Support Syst.* 150, 113664. doi: 10.1016/j.dss.2021.113664

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q.* 13, 319. doi: 10.2307/249008

Debitz, U., Hans-Eberhardt, P., and Richter, P. (2016). *BMS – Beanspruchungs-mess-skalen | Testzentrale.* Available online at: https://www.testzentrale.de/shop/beanspruchungs-mess-skalen.html (accessed June 28, 2023).

Dettmers, J., and Krause, A. (2020). Der Fragebogen zur Gefährdungsbeurteilung psychischer Belastungen (FGBU). *Z. Arbeits Organisationspsychol.* 64, 99–119. doi: 10.1026/0932-4089/a000318

Dietzmann, C., and Duan, Y. (2022). "Artificial intelligence for managerial information processing and decision- making in the era of information overload," in *Proceedings of the Annual Hawaii International Conference on System Sciences. Proceedings of the 55th Hawaii International Conference on System Sciences.*

Doshi-Velez, F., and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv* 1702.08608. doi: 10.48550/arXiv.1702.08608

Douglas, B. D., Ewell, P. J., and Brauer, M. (2023). Data quality in online human-subjects research: comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLoS ONE* 18, e0279720. doi: 10.1371/journal.pone.0279720

Gaube, S., Suresh, H., Raue, M., Lermer, E., Koch, T. K., Hudecek, M. F. C., et al. (2023). Non-task expert physicians benefit from correct explainable AI advice when reviewing X-rays. *Sci. Rep.* 13, 1383. doi: 10.1038/s41598-023-28633-w

Gimpel, H., Berger, M., Regal, C., Urbach, N., Kreilos, M., Becker, J., et al. (2020). *Belastungsfaktoren der Digitalen Arbeit: Eine beispielhafte Darstellung der Faktoren, die Digitalen Stress Hervorrufen.* Available online at: https://eref.uni-bayreuth.de/55149 (accessed June 28, 2023).

Glaser, J., Seubert, C., Hornung, S., and Herbig, B. (2015). The impact of learning demands, work-related resources, and job stressors on creative performance and health. *J. Pers. Psychol.* 14, 37–48. doi: 10.1027/1866-5888/a000127

Goodhue, D. L., and Thompson, R. L. (1995). Task-technology fit and individual performance. *MIS Q.* 19, 213. doi: 10.2307/249689

Gorla, N., Somers, T. M., and Wong, B. (2010). Organizational impact of system quality, information quality, and service quality. *J. Strateg. Inf. Syst.* 19, 207–228. doi: 10.1016/j.jsis.2010.05.001

Hacker, W. (1978). *Allgemeine Arbeits- und Ingenieurpsychologie: Psychische Struktur und Regulation von Arbeitstätigkeiten.* Schriften zur Arbeitspsychologie (Huber).

Hacker, W. (2016). Vernetzte künstliche Intelligenz/Internet der Dinge am deregulierten Arbeitsmarkt: Psychische Arbeitsanforderungen. *J. Psychol. Alltagshandelns.* 9, 4–21.

Haefner, N., Wincent, J., Parida, V., and Gassmann, O. (2021). Artificial intelligence and innovation management: a review, framework, and research agenda. *Technol. Forecasting Soc. Change* 162, 120392. doi: 10.1016/j.techfore.2020.120392

Hart, S. G., and Staveland, L. E. (1988). Development of NASA-TLX (task load index): results of empirical and theoretical research. *Adv. Psychol.* 52, 139–183. doi: 10.1016/S0166-4115(08)62386-9

Hassani, H., Silva, E. S., Unger, S., TajMazinani, M., and Mac Feely, S. (2020). Artificial intelligence (AI) or intelligence augmentation (IA): what is the future? *AI* 1, 143–155. doi: 10.3390/ai1020008

Heinisch, C. (2002). Inmitten der Informationsflut herrscht Informationsmangel. *ABI-Technik* 22, 340–349. doi: 10.1515/ABITECH.2002.22.4.340

Hellebrandt, T., Huebser, L., Adam, T., Heine, I., and Schmitt, R. H. (2021). Augmented intelligence – Mensch trifft Künstliche Intelligenz. *Zeitschrift für wirtschaftlichen Fabrikbetrieb* 116, 433–437. doi: 10.1515/zwf-2021-0104

Helquist, J. H., Deokar, A., Meservy, T., and Kruse, J. (2011). Dynamic collaboration. *SIGMIS Database* 42, 95–115. doi: 10.1145/1989098.1989104

Henkel, M., Horn, T., Leboutte, F., Trotsenko, P., Dugas, S. G., Sutter, S. U., et al. (2022). Initial experience with AI Pathway Companion: evaluation of dashboard-enhanced clinical decision making in prostate cancer screening. *PLoS ONE* 17, e0271183. doi: 10.1371/journal.pone.0271183

Hsiao, J. L., Wu, W. C., and Chen, R. F. (2013). Factors of accepting pain management decision support systems by nurse anesthetists. *BMC Med. Inform. Decis. Mak.* 13, 16. doi: 10.1186/1472-6947-13-16

Hu, L., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equ Model. Multidiscip. J.* 6, 1–55. doi: 10.1080/10705519909540118

Hussy, W., Schreier, M., and Echterhoff, G. (2013). *Forschungsmethoden in Psychologie und Sozialwissenschaften für Bachelor, 2nd Edn.* Berlin; Heidelberg: Springer.

Hwang, M. I., and Lin, J. W. (1999). Information dimension, information overload and decision quality. *J. Inf. Sci.* 25, 213–218. doi: 10.1177/016555159902500305

Iriani, S. S., and Andjarwati, A. L. (2020). Analysis of perceived usefulness, perceived ease of use, and perceived risk toward online shopping in the era of Covid-19 pandemic. *Syst. Rev. Pharm.* 11, 313–320.

ISO International Organization for Standardization (2017). *ISO 10075-1:2017. Ergonomic Principles Related to Mental Workload — Part 1: General Issues and Concepts, Terms and Definitions.* Available online at: https://www.iso.org/obp/ui/#iso:std:iso:10075:-1:ed-1:v1:en (accessed June 28, 2023).

ISO International Organization for Standardization (2019). *ISO 9241-210:2010. Ergonomics of Human-System Interaction: Human-Centred Design for Interactive Systems.* Available online at: https://www.iso.org/obp/ui/#iso:std:iso:9241:-210:ed-1:v1:en (accessed June 28, 2023).

ISO International Organization for Standardization (2020). *ISO 9241-210:2010. Ergonomics of Human-System Interaction: Human-Centred Design for Interactive Systems.* Available online at: https://www.iso.org/obp/ui/#iso:std:iso:9241:-210:ed-1:v1:en (accessed June 28, 2023).

Iwanowa, A. (2006). "Das Ressourcen-Anforderungen-Stressoren-Modell," in *Zur Psychologie der Tätigkeiten. Schriften zur Arbeitspsychologie*, eds. P. Sachse and W. Weber (Bern: Huber), 265–283.

Jacobs, M., Pradier, M. F., McCoy, T. H., Jr., Perlis, R. H., Doshi-Velez, F., and Gajos, K. Z. (2021). How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Transl. Psychiatry* 11, 108. doi: 10.1038/s41398-021-01224-x

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: human-AI symbiosis in organizational decision making. *Bus. Horiz.* 61, 577–586. doi: 10.1016/j.bushor.2018.03.007

Jin, M. H., and McDonald, B. (2017). Understanding employee engagement in the public sector: the role of immediate supervisor, perceived organizational support, and learning opportunities. *Am. Rev. Public Admin.* 47, 881–897. doi: 10.1177/0275074016643817

Jussupow, E., Spohrer, K., Heinzl, A., and Gawlitza, J. (2021). Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence. *Inform. Syst. Res.* 32, 713–735. doi: 10.1287/isre.2020.0980

Kalkbrenner, M. T. (2021). Enhancing assessment literacy in professional counseling: a practical overview of factor analysis. *Prof. Couns.* 11, 267–284. doi: 10.15241/mtk.11.3.267

Karasek, R. A. (1979). Job demands, job decision latitude, and mental strain: implications for job redesign. *Admin. Sci. Q.* 24, 285. doi: 10.2307/2392498

Kauffeld, S., and Schermuly, C. C. (2011). "Arbeitszufriedenheit und Arbeitsmotivation," in *Springer-Lehrbuch. Arbeits-, Organisations- und Personalpsychologie für Bachelor*, ed. S. Kauffeld (Berlin; Heidelberg: Springer), 179–194.

Kim, J., Davis, T., and Hong, L. (2022). "Augmented intelligence: enhancing human decision making," in *Bridging Human Intelligence and Artificial Intelligence*, eds. M. V. Albert, L. Lin, M. J. Spector, and L. S. Dunn (Cham: Springer), 151–170.

Kirste, M. (2019). "Augmented Intelligence – Wie Menschen mit KI zusammen Arbeiten," in *Open. Künstliche Intelligenz: Technologie, Anwendung, Gesellschaft, 1st Edn*, ed. V. Wittpahl (Berlin; Heidelberg: Springer Vieweg), 58–71.

Klopping, I. M., and McKinney, E. (2004). Extending the technology acceptance model and the task-technology fit model to consumer e-commerce. *Inf. Technol. Learn. Perform. J.* 22, 1.

Kohl, M. (2012). Performance measures in binary classification. *Int. J. Stat. Med. Res.* 1, 79–81. doi: 10.6000/1929-6029.2012.01.01.08

Koltay, T. (2017). "Information overload in a data-intensive world," in *Understanding Information*, ed. A. J. Schuster (Cham: Springer), 197–217.

Konys, A., and Nowak-Brzezińska, A. (2023). Knowledge engineering and data mining. *Electronics* 12, 927. doi: 10.3390/electronics12040927

Korner, S., and Brown, G. (1990). Exclusion of children from family psychotherapy: family therapists' beliefs and practices. *J. Fam. Psychol.* 3, 420–430. doi: 10.1037/h0080555

Körner, U., Müller-Thur, K., Lunau, T., Dragano, N., Angerer, P., and Buchner, A. (2019). Perceived stress in human-machine interaction in modern manufacturing environments-Results of a qualitative interview study. *Stress Health.* 35, 187–199. doi: 10.1002/smi.2853

Kraus, T., Gaschow, L., Eisenträger, M., and Wischmann, S. (2021). *Erklärbare KI – Anforderungen, Anwendungsfälle und Lösungen*. Available online at: https://vdivde-it.de/de/publikation/erklaerbare-ki-anforderungen-anwendungsfaelle-und-loesungen (accessed June 28, 2023).

Krcmar, H. (2011). *Einführung in das Informationsmanagement*. Berlin; Heidelberg: Springer-Lehrbuch. doi: 10.1007/978-3-642-15831-5

Krieger, P., and Lausberg, C. (2021). Entscheidungen, Entscheidungsfindung und Entscheidungsunterstützung in der Immobilienwirtschaft: Eine systematische Literaturübersicht. *Z. Immobilienökonomie.* 7, 1–33. doi: 10.1365/s41056-020-00044-2

Lackey, S. J., Salcedo, J. N., Szalma, J. L., and Hancock, P. A. (2016). The stress and workload of virtual reality training: the effects of presence, immersion and flow. *Ergonomics.* 59, 1060–1072. doi: 10.1080/00140139.2015.1122234

Langer, M., König, C. J., and Busch, V. (2021). Changing the means of managerial work: effects of automated decision support systems on personnel selection tasks. *J. Bus. Psychol.* 36, 751–769. doi: 10.1007/s10869-020-09711-6

Latos, B. A., Harlacher, M., Przybysz, P. M., and Mutze-Niewohner, S. (2017). "Transformation of working environments through digitalization: exploration and systematization of complexity drivers," in *IEEE International Conference on Industrial Engineering and Engineering Management (IEEM) 2017* (Singapore: IEEE Publications).

Lazarus, R. S., and Folkman, S. (1984). *Stress, Appraisal, and Coping.* New York, NY: Springer Publishing Company.

Lean, Y., and Shan, F. (2012). Brief review on physiological and biochemical evaluations of human mental workload. *Hum. Factors Man.* 22, 177–187. doi: 10.1002/hfm.20269

Lee, D., Moon, J., and Kim, Y. J. (2007). "The effect of simplicity and perceived control on perceived ease of use," in *AMCIS 2007 Proceedings*. Association for Information Systems.

Lee, M. H., Siewiorek, D. P., Smailagic, A., Bernardino, A., and Bermúdez i Badia, S. (2021). "A human-AI collaborative approach for clinical decision making on rehabilitation assessment," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM.

Li, D., Pehrson, L. M., Lauridsen, C. A., Tøttrup, L., Fraccaro, M., Elliott, D., et al. (2021). The added effect of artificial intelligence on physicians' performance in detecting thoracic pathologies on CT and chest X-ray: a systematic review. *Diagnostics*. 11, 2206. doi: 10.3390/diagnostics11122206

Lingmont, D. N. J., and Alexiou, A. (2020). The contingent effect of job automating technology awareness on perceived job insecurity: exploring the moderating role of organizational culture. *Technol. Forecasting Soc. Change*. 161, 120302. doi: 10.1016/j.techfore.2020.120302

Liu, C. F., Chen, Z. C., Kuo, S. C., and Lin, T. C. (2022). Does AI explainability affect physicians' intention to use AI? *Int. J. Med. Inform.* 168, 104884. doi: 10.1016/j.ijmedinf.2022.104884

Loukidou, L., Loan-Clarke, J., and Daniels, K. (2009). Boredom in the workplace: more than monotonous tasks. *Int. J. Manag. Rev.* 11, 381–405. doi: 10.1111/j.1468-2370.2009.00267.x

Lu, L., and Argyle, M. (1991). Happiness and cooperation. *Pers. Individ. Dif.* 12, 1019–1030. doi: 10.1016/0191-8869(91)90032-7

Lyell, D., Magrabi, F., and Coiera, E. (2018). The effect of cognitive load and task complexity on automation bias in electronic prescribing. *Hum. Factors*. 60, 1008–1021. doi: 10.1177/0018720818781224

Machdar, N. M. (2019). The effect of information quality on perceived usefulness and perceived ease of use. *Bus. Entrep. Rev.* 15, 131–146. doi: 10.25105/ber.v15i2.4630

Maes, P. (1995). "Agents that reduce work and information overload," in *Readings in Human–Computer Interaction* (Elsevier), 811–821.

Mayer, A.-S., Strich, F., and Fiedler, M. (2020). Unintended consequences of introducing AI systems for decision making. *MIS Q. Exec.* 19, 239–257. doi: 10.17705/2msqe.00036

Maynard, D. C., and Hakel, M. D. (1997). Effects of objective and subjective task complexity on performance. *Hum. Perform.* 10, 303–330. doi: 10.1207/s15327043hup1004_1

Mercado, J. E., Rupp, M. A., Chen, J. Y. C., Barnes, M. J., Barber, D., and Procci, K. (2016). Intelligent agent transparency in human-agent teaming for multi-UxV management. *Hum. Factors*. 58, 401–415. doi: 10.1177/0018720815621206

Meske, C., and Bunde, E. (2022). Design principles for user interfaces in AI-based decision support systems: the case of explainable hate speech detection. *Inf. Syst. Front.* 25, 743–773. doi: 10.1007/s10796-021-10234-5

Miller, D., and Lee, J. (2001). The people make the process: commitment to employees, decision making, and performance. *J. Manag.* 27, 163–189. doi: 10.1177/014920630102700203

Minge, M., Thüring, M., Wagner, I., and Kuhr, C. V. (2017). "The meCUE questionnaire: a modular tool for measuring user experience," in *Advances in Ergonomics Modeling, Usability and Special Populations*, eds. M. Soares, C. Falcão, and T. Z. Ahram (Cham: Springer), 115–128.

Misra, S., Roberts, P., and Rhodes, M. (2020). Information overload, stress, and emergency managerial thinking. *Int. J. Disaster Risk Reduc.* 51, 101762. doi: 10.1016/j.ijdrr.2020.101762

Mohr, G., Müller, A., Rigotti, T., Aycan, Z., and Tschan, F. (2006). The assessment of psychological strain in work contexts. *Eur. J. Psychol. Assess.* 22, 198–206. doi: 10.1027/1015-5759.22.3.198

Moosbrugger, H., and Kelava, A. (2008). *Testtheorie und Fragebogenkonstruktion*. Berlin; Heidelberg: Springer.

Morgeson, F. P., and Humphrey, S. E. (2006). The Work Design Questionnaire (WDQ): developing and validating a comprehensive measure for assessing job design and the nature of work. *J. Appl. Psychol.* 91, 1321–1339. doi: 10.1037/0021-9010.91.6.1321

Mosaly, P. R., Mazur, L. M., Yu, F., Guo, H., Derek, M., Laidlaw, D. H., et al. (2018). Relating task demand, mental effort and task difficulty with physicians' performance during interactions with electronic health records (EHRs). *Int. J. Hum. Comput. Interact.* 34, 467–475. doi: 10.1080/10447318.2017.1365459

Mvududu, N. H., and Sink, C. A. (2013). Factor analysis in counseling research and practice. *Couns. Outcome Res. Eval.* 4, 75–98. doi: 10.1177/2150137813494766

Nicodeme, C. (2020). "Build confidence and acceptance of AI-based decision support systems - Explainable and liable AI," in *13th International Conference on Human System Interaction (HSI) 2020*. Tokyo: IEEE Publications.

Nisar, S. K., and Rasheed, M. I. (2020). Stress and performance: investigating relationship between occupational stress, career satisfaction, and job performance of police employees. *J. Public Aff.* 20, e1986. doi: 10.1002/pa.1986

Omar, N., Munir, Z. A., Kaizan, F. Q., Noranee, S., and Malik, S. A. (2019). The impact of employees motivation, perceived usefulness and perceived ease of use on employee performance among selected public sector employees. *Int. J. Acad. Res. Bus. Soc. Sci.* 9, 6074. doi: 10.6007/IJARBSS/v9-i6/6074

Panigutti, C., Beretta, A., Giannotti, F., and Pedreschi, D. (2022). "Understanding the impact of explanations on advice-taking: a user study for AI-based clinical Decision Support Systems," in *CHI Conference on Human Factors in Computing System*, eds. S. Barbosa, C. Lampe, C. Appert, Shamma, A. David, S. Drucker, J. Williamson, and K. Yatani (New York, NY: Association for Computing Machinery), 1–9.

Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Ruggieri, S., and Turini, F. (2019). Meaningful explanations of black Box AI decision systems. *AAAI.* 33, 9780–9784. doi: 10.1609/aaai.v33i01.33019780

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., and Damer, E. (2022). Data quality of platforms and panels for online behavioral research. *Behav. Res. Methods*. 54, 1643–1662. doi: 10.3758/s13428-021-01694-3

Phillips-Wren, G., and Adya, M. (2020). Decision making under stress: the role of information overload, time pressure, complexity, and uncertainty. *J. Decis. Syst.* 29(Suppl. 1), 213–225. doi: 10.1080/12460125.2020.1768680

Porst, R. (2013). *Fragebogen: Ein Arbeitsbuch*. Wiesbaden: Springer.

Prakash, A. V., and Das, S. (2020). Intelligent conversational agents in mental healthcare services: a thematic analysis of user perceptions. *PAJAIS* 12, 1–34. doi: 10.17705/1thci.12201

Rajula, H. S. R., Verlato, G., Manchia, M., Antonucci, N., and Fanos, V. (2020). Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. *Medicina* 56, 455. doi: 10.3390/medicina56090455

Rau, R., Schweden, F., Hoppe, J., and Hacker, W. (2021). *Verfahren zur Tätigkeitsanalyse und-Gestaltung bei Mentalen Arbeitsanforderungen (TAG-MA)*. Kröning: Asanger.

Richter, G. (2000). *Psychische Belastung und Beanspruchung. Stress, psychische Ermüdung, Monotonie, psychische Sättigung*. Dortmund: Bundesamt für Arbeitsschutz und Arbeitsmedizin.

Richter, G., Henkel, D., Rau, R., and Schütte, M. (2014). "Infoteil A: Beschreibung psychischer Belastungsfaktoren bei der Arbeit," in *Erfahrungen und Empfehlungen*, eds. B. f.ü.r. Berlin: Arbeitsschutz, und Arbeitsmedizin (Chair), Gefährdungsbeurteilung psychischer Belastung.

Rigopoulos, G., Psarras, J., and Th. Askoun, D. (2008). A TAM model to evaluate user's attitude towards adoption of decision support systems. *J. Appl. Sci.* 8, 899–902. doi: 10.3923/jas.2008.899.902

Rohmert. (1984). Das Belastungs-Beanspruchungs-Konzept. Z. *Arbeitswissenschaft*. 38, 193–200.

Rowden, R. W., and Conine, C. T. (2005). The impact of workplace learning on job satisfaction in small US commercial banks. *J. Workplace Learn*. 17, 215–230. doi: 10.1108/13665620510597176

Rudolph, E., Schönfelder, E., and Hacker, W. (2017). *Tätigkeitsbewertungssystem – Geistige Arbeit (Langform) TBS-GA (L) (PT Verlag)* PT Verlag.

Rusli, B. N., Edimansyah, B. A., and Naing, L. (2008). Working conditions, self-perceived stress, anxiety, depression and quality of life: a structural equation modelling approach. *BMC Public Health* 8, 48. doi: 10.1186/1471-2458-8-48

Sailer, M. (2016). *Wirkung von Gamification auf Motivation*. Wiesbaden: Springer.

Sati, R. A. S., and Ramaditya, M. R. (2020). *Effect of Perception of Benefits, Easy Perception of Use, Trust and Risk Perception Towards Interest Using E-money*. Sekolah Tinggi Ilmu Ekonomi Indonesia Jakarta.

Saxena, D., and Lamest, M. (2018). Information overload and coping strategies in the big data context: evidence from the hospitality sector. *J. Inf. Sci.* 44, 287–297. doi: 10.1177/0165551517693712

Schulz-Dadaczynski, A. (2017). Umgang mit Zeit-und Leistungsdruck. *Eher Anpassung als Reduktion. Präv. Gesundheitsf.* 12, 160–166. doi: 10.1007/s11553-017-0582-5

Semmer, N. K., and Zapf, D. (2018). "Theorien der Stressentstehung und -bewältigung," in *Handbuch Stressregulation und Sport*, eds. R. Fuchs and M. Gerber (Berlin; Heidelberg: Springer), 23–50.

Shin, D. (2020). User perceptions of algorithmic decisions in the personalized AI system:perceptual evaluation of fairness, accountability, transparency, and explainability. *J. Broadcast. Electron. Media*. 64, 541–565. doi: 10.1080/08838151.2020.1843357

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: implications for explainable AI. *Int. J. Hum. Comput. Stud.* 146, 102551. doi: 10.1016/j.ijhcs.2020.102551

Shin, D., Zhong, B., and Biocca, F. A. (2020). Beyond user experience: what constitutes algorithmic experiences? *Int. J. Inform. Manag.* 52, 1–11. doi: 10.1016/j.ijinfomgt.2019.102061

Shrivastav, H., and Kongar, E. (2021). "Information overload in organization: impact on decision making and influencing strategies," in *IEEE Technology and Engineering Management Conference – Europe (TEMSCON-EUR) 2021*. Dubrovnik: IEEE Publications.

Silva, J. P., and Gonçalves, J. (2022). Process standardization: "The driving factor for bringing artificial intelligence and management analytics to SMEs," in *10th International Symposium on Digital Forensics and Security (ISDFS) 2022*. Istanbul: IEEE Publications.

Skitka, L. J., Mosier, K. L., and Burdick, M. (1999). Does automation bias decision-making? *Int. J. Hum. Comput. Stud.* 51, 991–1006. doi: 10.1006/ijhc.1999.0252

Spector, J. M., and Ma, S. (2019). Inquiry and critical thinking skills for the next generation: from artificial intelligence back to human intelligence. *Smart Learn. Environ.* 6, 1–11. doi: 10.1186/s40561-019-0088-z

Steil, A. V., de Cuffa, D., Iwaya, G. H., and Pacheco, R., Cd,. S. (2020). Perceived learning opportunities, behavioral intentions and employee retention in technology organizations. *J. Workplace Learn*. 32, 147–159. doi: 10.1108/JWL-04-2019-0045

Stemmann, J., and Lang, M. (2014). Theoretische Konzeption einer allgemeinen technischen Problemlösefähigkeit und Möglichkeiten ihrer Diagnose. *J. Tech. Educ.* 2, 80–101.

Stenzl, A., Sternberg, C. N., Ghith, J., Serfass, L., Schijvenaars, B. J. A., and Sboner, A. (2022). Application of artificial intelligence to overcome clinical information overload in urological cancer. *BJU Int.* 130, 291–300. doi: 10.1111/bju.15662

Stowers, K., Kasdaglis, N., Rupp, M. A., Newton, O. B., Chen, J. Y. C., and Barnes, M. J. (2020). The Impact of agent transparency on human performance. *IEEE Trans. Hum. Mach. Syst.* 50, 245–253. doi: 10.1109/THMS.2020.2978041

Suki, N. M., and Suki, N. M. (2011). Exploring the relationship between perceived usefulness, perceived ease of use, perceived enjoyment, attitude and subscribers' intention towards using 3G mobile services. *J. Inf. Technol. Manag.* 22, 1–7.

Syrek, C. J., Apostel, E., and Antoni, C. H. (2013). Stress in highly demanding IT jobs: transformational leadership moderates the impact of time pressure on exhaustion and work-life balance. *J. Occup. Health Psychol.* 18, 252–261. doi: 10.1037/a0033085

Theron, J. C. (2014). Dying for information? An investigation into the effects of information overload in the UK and worldwide. *S. Afr. J. Libr. Inf. Sci.* 66, 1454. doi: 10.7553/66-1-1454

Timiliotis, J., Blümke, B., Serfözö, P. D., Gilbert, S., Ondrésik, M., Türk, E., et al. (2022). A novel diagnostic decision support system for medical professionals: prospective feasibility study. *JMIR Form. Res.* 6, e29943. doi: 10.2196/29943

Trägner, U. (2006). *Arbeitszeitschutzrechtliche Bewertung der Intensität von Arbeitsleistungen: Unter Besonderer Berücksichtigung der Rechtsprechung des Europäischen Gerichtshofes zum Bereitschaftsdienst*. Hartung-Gorre Verlag.

Tsakonas, G., and Papatheodorou, C. (2008). Exploring usefulness and usability in the evaluation of open access digital libraries. *Inf. Process. Manag.* 44, 1234–1250. doi: 10.1016/j.ipm.2007.07.008

Tutun, S., Johnson, M. E., Ahmed, A., Albizri, A., Irgil, S., Yesilkaya, I., et al. (2023). An AI-based decision support system for predicting mental health disorders. *Inf. Syst. Front.* 25, 1261–1276. doi: 10.1007/s10796-022-10282-5

Ulich, E. (2011). *Arbeitspsychologie*. Zurich: Vdf Hochschulverlag.

van Laar, E., van Deursen, A. J. A. M., van Dijk, J. A. G. M., and de Haan, J. (2017). The relation between 21st-century skills and digital skills: a systematic literature review. *Comput. Hum. Behav.* 72, 577–588. doi: 10.1016/j.chb.2017.03.010

Venkatesh, V., and Davis, F. D. (2000). A theoretical extension of the technology acceptance model: four longitudinal field studies. *Manag. Sci.* 46, 186–204. doi: 10.1287/mnsc.46.2.186.11926

Vosloban, R. I. (2012). The influence of the employee's performance on the Company's growth - A managerial perspective. *Procedia Econ. Fin.* 3, 660–665. doi: 10.1016/S2212-5671(12)00211-0

Walch, K. (2020). *Is There A Difference Between Assisted Intelligence Vs. Augmented Intelligence?* Forbes. Available online at: https://www.forbes.com/sites/cognitiveworld/2020/01/12/is-there-a-difference-between-assisted-intelligence-vs-augmented-intelligence/?sh=a23756f26aba (accessed July 7, 2023).

Wanner, J. (2021). Do you really want to know why? "Effects of AI-based DSS advice on human decisions," in *Proceedings of Americas Conference on Information Systems (AMCIS) 2021*.

Westenberger, J., Schuler, K., and Schlegel, D. (2022). Failure of AI projects: understanding the critical factors. *Procedia Comput. Sci.* 196, 69–76. doi: 10.1016/j.procs.2021.11.074

Wilkens, U. (2020). Artificial intelligence in the workplace – A double-edged sword. *Int. J. Inf. Learn. Technol.* 37, 253–265. doi: 10.1108/IJILT-02-2020-0022

Wook Seo, Y., Chang Lee, K., and Sung Lee, D. (2013). The impact of ubiquitous decision support systems on decision quality through individual absorptive capacity and perceived usefulness. *Online Inf. Rev.* 37, 101–113. doi: 10.1108/14684521311311658

Yin, J., and Qiu, X. (2021). Ai technology and online purchase intention: structural equation model based on perceived value. *Sustainability*. 13, 5671. doi: 10.3390/su13105671

Zapf, D. (1998). *Psychische Belastungen in der Arbeitswelt – Ein Überblick*. Available online at: https://www.guss-net.de/fileadmin/media/Projektwebsites/Guss-Net/Dokumente/service/downloads/allgemeine_infos_arbeit_gesundheit/08_Psychische_Belastungen_09.pdf (accessed June 28, 2023).

Zeffane, R., and McLoughlin, D. (2006). Cooperation and stress: exploring the differential impact of job satisfaction, communication and culture. *Manag. Res. Rev.* 29, 618–631. doi: 10.1108/01409170610712326

Zhang, S., Zang, X., and Zhang, F. (2021). Development and validation of the win-win scale. *Front. Psychol.* 12, 657015. doi: 10.3389/fpsyg.2021.657015

Zhang, Y., Liao, Q. V., and Bellamy, R. K. E. (2020). "Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making." in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (New York, NY: Association for Computing Machinery), 295–305.

frontiers | Frontiers in Artificial Intelligence

TYPE Perspective
PUBLISHED 27 September 2023
DOI 10.3389/frai.2023.1252897

OPEN ACCESS

EDITED BY
Uta Wilkens,
Ruhr-University Bochum, Germany

REVIEWED BY
Riccardo De Benedictis,
National Research Council (CNR), Italy
Eva Bittner,
University of Hamburg, Germany

*CORRESPONDENCE
Vera Hagemann
✉ vhagemann@uni-bremen.de

RECEIVED 04 July 2023
ACCEPTED 04 September 2023
PUBLISHED 27 September 2023

CITATION
Hagemann V, Rieth M, Suresh A and Kirchner F (2023) Human-AI teams—Challenges for a team-centered AI at work.
Front. Artif. Intell. 6:1252897.
doi: 10.3389/frai.2023.1252897

Human-AI teams—Challenges for a team-centered AI at work

Vera Hagemann[1]*, Michèle Rieth[1], Amrita Suresh[2] and Frank Kirchner[2,3]

[1]Business Psychology and Human Resources, Faculty of Business Studies and Economics, University of Bremen, Bremen, Germany, [2]Robotics Research Group, Faculty of Mathematics and Computer Science, University of Bremen, Bremen, Germany, [3]DFKI GmbH, Robotics Innovation Center, Bremen, Germany

As part of the Special Issue topic "Human-Centered AI at Work: Common Ground in Theories and Methods," we present a perspective article that looks at human-AI teamwork from a team-centered AI perspective, i. e., we highlight important design aspects that the technology needs to fulfill in order to be accepted by humans and to be fully utilized in the role of a team member in teamwork. Drawing from the model of an idealized teamwork process, we discuss the teamwork requirements for successful human-AI teaming in interdependent and complex work domains, including e.g., responsiveness, situation awareness, and flexible decision-making. We emphasize the need for team-centered AI that aligns goals, communication, and decision making with humans, and outline the requirements for such team-centered AI from a technical perspective, such as cognitive competence, reinforcement learning, and semantic communication. In doing so, we highlight the challenges and open questions associated with its implementation that need to be solved in order to enable effective human-AI teaming.

KEYWORDS

human-agent teaming, hybrid multi-team systems, cooperation, communication, teamwork, integrative artificial intelligence

1. Introduction

In the future, Mars is a target for long-duration space missions (Salas et al., 2015). Both governments and private space industries are fascinated by the Red Planet, and are aiming to send teams of astronauts on a mission to Mars in the late 2030's (Buchanan, 2017; NASA, 2017). For successful survival and operation on Mars, a habitat with intelligent systems, such as integrative Artificial Intelligence (Kirchner, 2020), and robots (e.g., for outdoor operations), are indispensable, among other things. To avoid unnecessary exposure to radiation the crew will be in the habitat most of the time. There, they will collaborate with technical systems with capabilities that are more like the cognitive abilities of humans compared to previous support systems. Advancements in Machine Learning and Artificial Intelligence (AI) have led to the development of systems that can handle uncertainties, adjust to changing situations, and make intelligent decisions independently (O'Neill et al., 2022). Intelligent autonomous agents can either exist as virtual entities or can embody a physical system such as a robot. Although much of the decision-making paradigm may be similar in both cases, the physical spatio-temporal constraints of robots must be properly considered in their decisions (Kabir et al., 2019). In the given context, autonomous agents perform tasks such as adaptively controlling light, temperature, and oxygen levels. In addition, they can gather important information about the outdoor environment and guide the crew's task planning by telling them when, for example, an outdoor mission is most advantageous due to weather conditions such as isotope storms. Additionally, for outdoor activities, multi-robot

teams (Cordes et al., 2010) will facilitate efficient exploration in areas of low accessibility, transportation of materials, and analysis and transmission of information to the human crew.

In the previously described scenario, we are concerned with human-AI teaming (cf. Schecter et al., 2022), also often referred to as human-agent teaming (cf. Schneider et al., 2021) or human autonomy teaming (cf. O'Neill et al., 2022) (all abbreviated as HAT). Those are systems, in which humans and intelligent, autonomous agents work interdependently toward common goals (O'Neill et al., 2022). These forms of hybrid teamwork (cf. Schwartz et al., 2016a,b) are already present in some industries and workplaces and are becoming more and more relevant, for example in aviation, civil protection, firefighting or medicine. They provide opportunities for increased safety at work and productivity, thus supporting human and organizational performance.

A well-known example from the International Space Station is the astronaut assistant CIMON-2 (Crew Interactive MObile companioN), which has already worked with the astronauts. CIMON-2 is controlled by voice and aims to support astronauts primarily in their workload of experiments, maintenance, and repair work. Astronauts can also activate linguistic emotion analysis, so that the agent can respond empathically to its conversation partners (DLR, 2020). Another example of AI at work is the chatbot CARL (Cognitive Advisor for Interactive User Relationship and Continuous Learning), which has been in use in the human resources department of Siemens AG. CARL can provide information on a wide range of human resources topics and thus serves as a direct point of contact for all employees. Also, the human resources Shared Service Experts themselves use CARL as a source of information in their work. CARL understands, advises, and guides and is used extensively within the company. Carl has been positively received by the employees as well as the human resources experts and leads to a facilitation in the work like a colleague (IBM and ver.di, 2020). Artificial agents are also used in the medical sector, for example, when nurses and robots collaborate efficiently in the Emergency Department during high workload situations, such as resuscitation or surgery.

The question that emerges is how to effectively design such a novel form of teamwork that fully meets the needs of humans in a successful teamwork process (Seeber et al., 2020; Rieth and Hagemann, 2022). This perspective thus highlights the role of AI interacting with humans in a team instead of only using a normal high developed technology. Consequently, this perspective aims toward a comprehensive and interdisciplinary exploration of the key factors that contribute to successful collaboration between humans and AI. We (1) illustrate the requirements for successful teamwork in interdependent and complex work domains based on the model of an idealized teamwork process, (2) identify the implications of these requirements for successful human-AI teaming, and (3) outline the requirements for AI to be team-centered from a technical perspective. Our goal is to draw attention to the teamwork-related requirements to enable effective human-AI teaming, also in hybrid multi-team systems, and at the same time to create awareness of what this means for the design of technology.

2. Human-AI multi-team systems

As described above, the question arises as to what aspects need to be considered in human-centered AI in teamwork, both in terms of the human crew and the "team" of artificial agents to achieve effective and safe team performance. Imagine a scenario for major disasters on earth. Here we will not only have a team with one or two humans and one agent, but several teams of people, e.g., police, fire fighters, rescue services, and several agents, e.g., assistance system in the control center, robots in buildings and drones in the air, who must communicate and collaborate successfully.

Thus, HATs also exist in a larger context and work in dependence with other teams. These human-AI multi teams are called hybrid multi-team systems (HMTS) and refer to "multiple teams consisting of n-number of humans and n-number of semi-autonomous agents [i.e., AI] having interdependence relationships with each other" (Schraagen et al., 2022, p. 202). They consist of sub-teams, with each individual and team striving to achieve hierarchically structured goals. Lower-level goals require coordination processes within a single team and higher-level goals require coordination with other teams. Their interaction is shaped by the varying degrees of task interdependencies between the sub-teams (Zaccaro et al., 2012). HMTS highlight the complexity of the overall teamwork situation, as sub-teams consist of humans, of agents and of humans and agents. Therefore, teamwork relevant constructs such as communication (Salas et al., 2005), building and maintaining an effective situation awareness (Endsley, 1999) and shared mental models (Mathieu et al., 2000) as well as decision making (Waller et al., 2004) will not only be of high relevance in the human crew, but also in the AI teams (Schwartz et al., 2016b) as well as in the human-AI teams (cf. e.g., Carter-Browne et al., 2021; Stowers et al., 2021; National Academies of Sciences, Engineering, and Medicine, 2022; O'Neill et al., 2022; Rieth and Hagemann, 2022).

To date, research has focused on individual facets of successful HAT, ignoring the Input-Process-Output (IPO) framework (Hackman, 1987) in teamwork research (cf. O'Neill et al., 2023) which acknowledges the pivotal role of group processes (e.g., shared mental models or communication) in converting inputs (e.g., autonomy or task) into desired outcomes (e.g., team performance or work satisfaction). Often, the focus is on single aspects such as trust (Lyons et al., 2022), agent autonomy (Ulfert et al., 2022), shared mental models (Andrews et al., 2023), or speech (Bogg et al., 2021). Examining individual facets is important to understand human-AI teaming, yet we would like to point out that *successful teamwork* does not consist of individual components *per se*, but rather the big picture, i.e., *the interaction of inputs, processes and outcomes*. Thus, we would like to think of a *teamwork-centered AI holistically* and *discuss relevant aspects for a successful human-AI teaming from a psychological and technical perspective using the model of the idealized teamwork process* (Hagemann and Kluge, 2017; see the black elements in Figure 1).

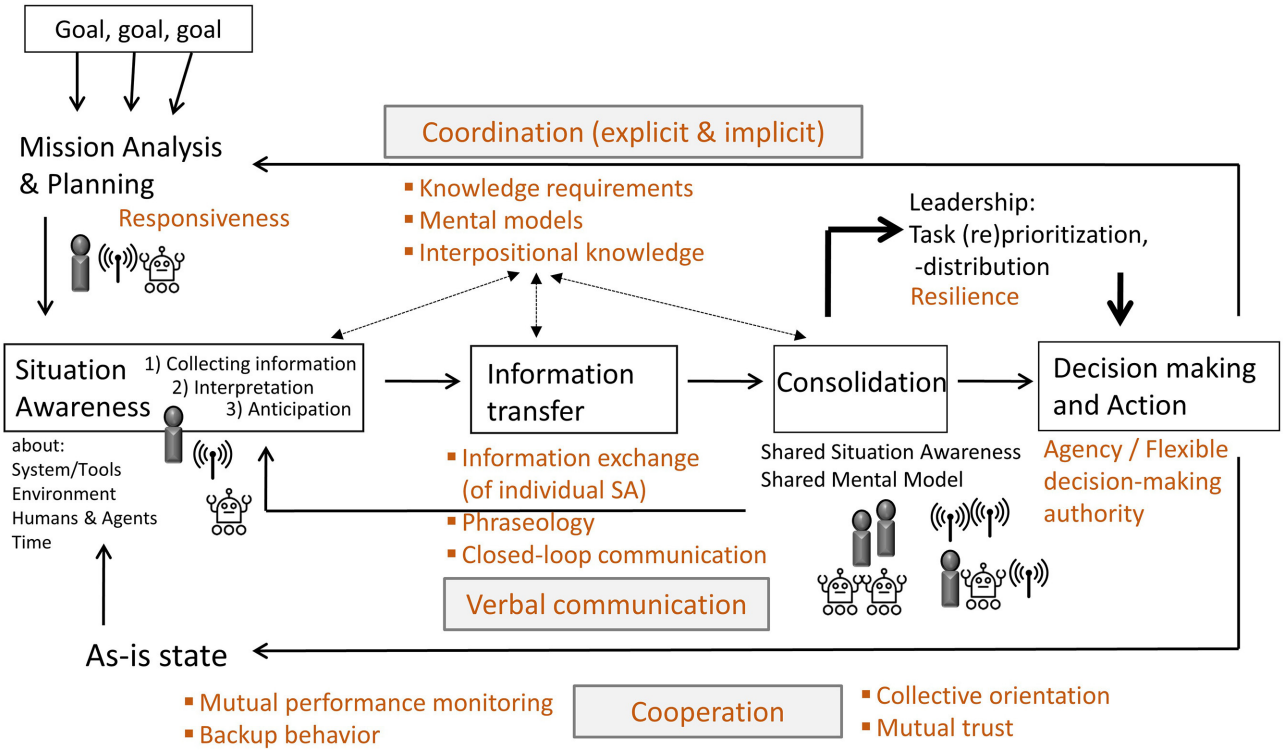

FIGURE 1
Model of an idealized teamwork process of hybrid multi-team systems. Adapted from Hagemann and Kluge (2017), licensed under CC BY 4.0.

2.1. Teamwork requirements in human-AI teaming

The cognitive requirements for effective teamwork and the team process demands are consolidated within the model of an idealized teamwork process (Kluge et al., 2014; Hagemann and Kluge, 2017). Figure 1 shows an adapted version of this model. The black elements in the model are from the original model by Hagemann and Kluge (2017). The brown elements of the model are additions which, based on literature analyses, are essential for team-centered AI addressing human needs and thus for successful human-AI teaming. These elements will now be discussed in more detail in the course of this article. Following the IPO model, our proposed model does not focus on solely a certain component of the model, such as only the input, but holistically all three components. Central elements of the model are situation awareness, information transfer, consolidation of individual mental models, leadership, and decision making (for a detailed explanation see Hagemann and Kluge, 2017). Human-AI teams are responsible for reaching specific goals (see top left of model), for example, search for, transport, and care for injured persons during a large-scale emergency, as well as extinguishing fires. Based on the overall goals, various sub goals exist for the all-human teams, the agent teams, and the human-AI teams that will be identified at the beginning of the teamwork process and communicated within the HMTS. For routine situations, there will be standard operating procedures known by all humans and agents. However, it becomes challenging for novel or unforeseen situations for which standard operating procedures do not yet exist. Here, an effective start requires an intensive exchange of mission analysis, goal specification, and strategy formulation, which are important teamwork processes occurring during planning activities (cf. Marks et al., 2001). Such planning activities are a major challenge, especially in multi-team systems, since between-team coordination is more difficult to achieve than within-team coordination, but it is also more important for effective multi-team system teamwork (Schraagen et al., 2022). Thus, the *responsiveness* of the agents will be important for a team-centered AI (see upper left corner in the model), meaning that the *agents are able to align their goals and interaction strategies to the shifting goals and intentions of others as well as the environment* (Lyons et al., 2022).

As depicted in our model, the defined goals provide the starting position for all teams building an effective situation awareness, which is important for successful collaboration within teams (Endsley, 1999; Flin et al., 2008). Situation awareness means collecting information from systems, tools, humans, agents, and environments, interpreting this information and anticipating future states. The continuous assessment of situations by all humans and agents is important, as they work independently as well as interdependently and each team needs to achieve a correct situation awareness and to share it within the HMTS. High-performing teams have been shown to spend more time sharing information and less time deciding on a plan, for example (Uitdewilligen and Waller, 2018). This implies the importance of a *sound and comprehensive situation awareness between humans and agents* (cf. McNeese et al., 2021b) and an accompanying

goal-oriented and continuous exchange of information in HMTS. For developing a shared situation awareness, the information transfer focuses on sending and receiving single situation awareness between team members. *Aligned phraseology* between humans and agents (i.e. using shared language and terminology) and *closed-loop communication* (i.e. verifying accurate message understanding through feedback: statement, repetition, reconfirmation; Salas et al., 2005) are essential for effective teamwork. However, possible effects of closed-loop communication have not yet been investigated in a HAT. Thus, it is not clear, for example, whether this form of communication is more likely to be considered disruptive in joint work and whether it should be used only in specific situations, such as when performing particularly important or sensitive tasks. Nevertheless, these requirements for communication in a HAT are important to consider for successful teamwork, as it has been shown that performance and perception of teamwork are significantly higher with verbal communication in a HAT (Bogg et al., 2021). Therefore, for a team-centered AI, *agents should communicate quite naturally with human team members in verbal language*.

Expectations of all humans and agents based on their mental models and interpositional knowledge impact the situation awareness, the information transfer, and the consolidation phases. Mental models are cognitive representations of system states, tasks, and processes, for example, and help humans and agents to describe, explain, and predict situations (Mathieu et al., 2000). Interpositional knowledge refers to an understanding of the tasks and needs of all team members to develop an understanding of the impact of one's actions on the actions of other team members and vice versa. It lays a foundation for understanding the information needs of others and the assistance they require (Smith-Jentsch et al., 2001). Interpositional knowledge and mental models are important prerequisites for effective coordination in HMTS, i.e., temporally and spatially appropriately orchestrated actions (Andrews et al., 2023). Thus, a *fully comprehensive and up-to-date mental model of the agents about the tasks and needs of the other human and artificial team members is highly relevant for team-centered AI*.

Based on effective information transfer, a common understanding of tasks, tools, procedures, and competencies of all team members is developed in the consolidation phase in terms of shared mental models. These shared knowledge structures help teams adapt quickly to changes during high workload situations (Waller et al., 2004) and increase their performance (Mathieu et al., 2000). The advantage of shared mental models is that HMTS can shift from time-consuming explicit coordination to implicit coordination in such situations (cf. Schneider et al., 2021). For example, observable behaviors or explicit statements may cause the agent to exhibit appropriate behavior, such as a robot observes that the human has reached a certain point in the experiment and is already preparing the materials that the human will need in the next step. Accordingly, a team-centered AI must be able to *coordinate with the humans in the team not only explicitly, but also implicitly*. In addition, *the agents in a HMTS must be able to detect when there is a breakdown in collaboration between humans and agents, or between the different agents*, and intervene so that they can explicitly coordinate again.

As a result of the consolidation phase, the HMTS or leading humans and agents need to make decisions to take actions. Thus, it is important that the artificial agents have *agency*, i.e., they can have control over their actions and the decision authority to execute these actions (Lyons et al., 2022). For an effective collaboration of humans and agents, the HMTS needs a *flexible decision-making authority*, that is, authority dynamically shifting among the humans and agents in response to complex and changing situations (Calhoun, 2022; Schraagen et al., 2022). Requirements in this phase include task prioritization and distribution as well as re-prioritization and distribution of tasks according to changes in the situation or plan (Waller et al., 2004). The *resilience* of the system is thus also important for team-centered AI, so that the agents can adapt to changing processes and tasks (Lyons et al., 2022). In this phase, it is very important that the *agents can interpret the statements of all the others and continue to think about the situation together with the humans*. Only in this way can HMTS be as successful as only human high-performing teams. That is due to the fact that in the decision-making phase high-performing teams compared to low performing teams use more interpretation-interpretation sharing sequences: the process involves an initial statement made by one human or agent, followed by an interpretative response from another agent, leading to a subsequent statement by the first agent that builds upon and expands the reasoning and thus build a collective sensemaking (Uitdewilligen and Waller, 2018).

The result of decision-making and action flows back into individual situation awareness and the original goals are compared with the as-is state achieved. This model of a continuously idealized teamwork process includes diverse feedback loops that enable a HMTS to adapt to changing environments and goals. For the described processes to be successfully completed, cooperation is required within the HMTS. This includes, for example, *mutual performance monitoring*, in which humans and agents keep track of each other while performing their own tasks to detect and prevent possible mistakes at an early stage (Paoletti et al., 2021). Cooperation also requires *backup behavior* in the team, i.e., the discretionary help from other human or artificial team members as well as a distinct *collective orientation* of all members (Salas et al., 2005; Hagemann et al., 2021; Paoletti et al., 2021). For team-centered AI, the agents must be able to provide this support behavior for the other team members. A successful pass through the teamwork process model also depends on the *trust* of each team member (Hagemann and Kluge, 2017; McNeese et al., 2021a). Important for the trust of humans in agents is a reliable performance, i.e., as few to no errors as possible (Hoff and Bashir, 2015; Lyons et al., 2022). Nevertheless, the agent should not only be particularly reliable, but for a team-centered AI it should also be able *to turn to all members of the HMTS in new situations and request an exchange because it cannot get on by itself*.

2.2. Technical requirement for artificial team members

Increased autonomy enables agents to *make decisions independently* in different situations, i.e., to develop *situational awareness*, even in situations where there is only a limited possibility of human intervention. For agents to be part of HMTS, it is mandatory to achieve a level of *cognitive competence* that

allows them to grasp the intentions of their teammates (Demiris, 2007; Trick et al., 2019). This claim is much easier said than done as it requires the existence of mental models in agents that are comparable to the models that humans rely on, especially if they exchange information. However, such models cannot just be preprogrammed and then implanted into agents. One reason for this is that the process by which mental models are created in humans is still a subject under investigation (Westbrook, 2006; Tabrez et al., 2020). Even though this cognitive competence is required for team-centricity, this process is difficult to reproduce artificially. On the other hand, there is usually not just a single isolated model (or brain process) that generates human behavior, but rather an ensemble of models that are active at any given time and influence the observable outcome. Compared to humans that function based on cognitive decision-making, intuition, etc., machines are digitized, and act based on experience, their understanding of the current situation, and prediction models. These models must improve over time, based on a limited set of prerecorded data to move toward more accurate, robust systems. Thus, for future developments in HATs it would be important to design models with higher *predictive power*, which we define by how well the model can predict the outcome of its decisions based on the situation, experience and team behavior (see also Raileanu et al., 2018).

Moreover, agents must be competent and empowered to make decisions when needed, without having to wait for instructions from humans, especially in extreme environments where humans must adapt to particular conditions (Hambuchen et al., 2021), and resources are scarce. System *resilience* is also of high importance as the consequences of failure, on either side (human or agent) could be catastrophic. Whenever there is a potential threat to human lives, HMTS can prove more effective compared to homogeneous human or AI teams. During search and rescue operations on Earth (Govindarajan et al., 2016), *responsiveness, coordination, and effective communication* are crucial requirements for HMTS. Therefore, through teleoperation and on-site collaboration, HATs are able to mitigate the impact after a disaster. HATs can also be witnessed in modern medical applications that demand cooperation and high degrees of precision. For example, nurses and robots in the Emergency Department can efficiently handle high workloads, and safety-critical procedures like surgery and resuscitation, using a new *reinforcement learning* system design (Taylor, 2021). Reinforcement learning is a class of machine learning algorithms, wherein the agent receives either a reward or a penalty depending on the favorability of the outcome of a particular action.

In HATs of the future, we will thus have to work with agents that can learn over time to adjust to human behavior and shape the models of the environment and of other team members over time. This learning approach will enable the agents to exchange substantial information even with very few bits or in other words content and meaning will be exchangeable between humans and agents rather than bits and bytes. This process, also known as *semantic communication,* is currently under investigation by different teams from a more information theoretic approach over *symbolic reasoning* to an approach that is called *integrative artificial intelligence* (Kirchner, 2020). Beck et al. (2023) approach this problem by modeling semantic information as hidden random variables to achieve reliable communication under limited resources. This is a valuable step toward adapting to the problem of communication losses and latencies in applications like space, and exploration in remote areas. In a HAT setup, it is important to make some decisions regarding the nature of the team, either a priori or dynamically. Like pure human systems, assigning specific roles and defining hierarchies among agents in a team and between teams can enhance the overall mission strategy. Role-based task allocation is especially useful when the team consists of heterogeneous (Dettmann et al., 2022) and (or) reconfigurable (Roehr et al., 2014) agents. In HMTS, having every member trying to communicate with every other member is highly impractical, resource intensive and chaotic. This issue is further complicated when all members are authorized to act as they will. Implementing an *organized hierarchical team structure* (Vezhnevets et al., 2017) is therefore imperative for a team-centered AI successfully collaborating with humans.

To achieve seamless interaction between humans and agents, the latter must display behavioral traits that are acceptable to humans. An agent is truly team-centered when it can intelligently adapt to the situation and team requirements, in a *team-oriented* (Salas et al., 2005), rather than a dominant or submissive manner. Agents need to achieve predictive capabilities for other teammates and the environment to account for variation, as in Raileanu et al. (2018). In the autonomous vehicle domain, it is crucial that the vehicle can accurately predict the behavior of pedestrians and others to enable seamless navigation (Rhinehart et al., 2019). According to Teahan (2010), behavior is defined by how an agent acts while interacting with its environment. Interaction entails *communication* which can be either verbal or non-verbal. An interesting aspect will be to investigate deeper into large language models, like ChatGPT and to find out if these approaches can be extended to general interactive behavior (Park et al., 2023) instead of just text and images. Apart from language, verbal communication is also characterized by the acoustics of the voice, and style of speech. Moreover, movement is a fundamental component that defines the behavior of any team member. Depending on design, agents are already capable of performing and recognizing gestures (Wang et al., 2019; Xia et al., 2019) and emotions (Arriaga et al., 2017). Motion analyses have shown that the intention behind performing an action is intrinsically embedded in the style of movement, for instance, in the dynamics of the arm (Niewiadomski et al., 2021; Gutzeit and Kirchner, 2022). In all the scenarios, one of the biggest challenges faced by HMTS is the *trustworthiness* of the team. Cooperation requires building trust-based relationships between the team members. Bazela and Graczak (2023) evaluated, among other factors, "*the team's willingness to consider it* [the Kalman autonomous rover—an astronaut assistant] *a partially conscious team member*" (p. 369). The opposite also holds true. Agents must maintain a *high trust factor* of their human teammates, i.e., be able to trust humans, as this factor has a significant influence on the decision-making process (Chen et al., 2018). For instance, the agent's trust factor can be improved by means of reliable communication when the human switches strategies. From a technical perspective, humans are the chaos factor in the HAT equation and even though this

can be modeled to a certain extent on the agent side, an effective collaboration largely depends on the predictability of human actions. A summary of all the requirements mentioned for team-centered successful human-AI teaming addressing human needs can be found with definitions of these and example references in the Supplementary Table 1.

3. Conclusion

The aim of this contribution is to discuss the central teamwork facets for successful HATs in an interdisciplinary way. Starting from a psychological perspective addressing the human needs, the importance of team-centered AI is revealed. However, its technical feasibility is challenging. It is an open problem from a standpoint of technical cognition if AI systems can ever be regarded and/or accepted as actual team members as this poses a very fundamental question of AI. This question refers to the challenge to replicate intelligence in technical systems as we only know it from human systems. This is an old and long-standing question that has been addressed by Turing (1948) in his famous paper on "*intelligent machinery*" already in the early last century. As a mathematician he concluded that it is not possible to build such systems *ad hoc*. One loophole that he identified in this paper is to create highly articulated robotic systems that learn—in an *open-ended process*- from the inter*action* with a *real*-world environment. It is his assumption that somewhere along this process, which is *open-ended*, some of the features that we associate with intelligence may emerge and thus the resulting system will eventually be able to simulate intelligence well enough such that it will be regarded as intelligent by humans. If this is actually feasible has never been tested but could be a worthwhile experiment to perform with technologies of the 21st century. However, for the meantime, teamwork attributes like *responsiveness, situation awareness, closed-loop communication, mental models, and decision making* remain to be buzz words in this context and are technical features that we will be able to implement to a limited extent into technical systems in order to enable these systems to act as valuable tools for humans in well-defined environments and contexts. But whether this will qualify the agents as team members is unknown so far (cf. also Rieth and Hagemann, 2022). This would in fact require a much deeper understanding of the processes that enable cognition in human systems as we have it today and even if we had that understanding, it will still be an open question if the understanding of mental processes is also a blueprint or design approach to achieve the same in technical systems. Overall, the manuscript provides insights into the team-centered requirements for effective collaboration in HATs and underscores the importance of considering teamwork-related factors in the design of technology. Our proposed guidelines can be used to design and evaluate future concrete interactive systems. In the experimental testing of the single facets discussed for a truly team-centered and successful HAT, which considers the needs of the humans in the HAT, many highly specific further research questions will arise, the scientific treatment of which will be of great importance for the implementation of future HATs. Thus, further research in this area is needed to address the challenges and unanswered questions associated with HMTS. Solving them will open doors to applying hybrid systems in diverse setups, thus leveraging the advantages of both, human and agent members, as human-AI multi-team systems.

Author contributions

VH had the idea for this perspective article and took the lead in writing it. MR, AS, and FK contributed to the discussions and writing of this paper. All authors contributed to the writing and review of the manuscript and approved the version submitted.

Conflict of interest

FK was employed by DFKI GmbH.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2023.1252897/full#supplementary-material

References

Andrews, R. W., Lilly, J. M., Srivastava, D., and Feigh, K. M. (2023). The role of shared mental models in human-AI teams: a theoretical review. *Theor. Iss. Ergon. Sci.* 24, 129–175. doi: 10.1080/1463922X.2022.2061080

Arriaga, O., Valdenegro-Toro, M., and Plöger, P. (2017). Real-time convolutional neural networks for emotion and gender classification. *arXiv preprint*. 1710, 07557. doi: 10.48550/arXiv.1710.07557

Bazela, N., and Graczak, P. (2023). "HRI in the AGH space systems planetary rover team: A study of long-term human-robot cooperation," in *Social Robots in Social Institutions: Proceedings of Robophilosophy* eds R. Hakli, P. Mäkelä, and J. Seibt (Amsterdam: IOS Press), 361–370.

Beck, E., Bockelmann, C., and Dekorsy, A. (2023). Semantic information recovery in wireless networks. *arXiv preprint* 2204, 13366. doi: 10.3390/s23146347

Bogg, A., Birrell, S., Bromfield, M. A., and Parkes, A. M. (2021). Can we talk? How a talking agent can improve human autonomy team performance. *Theoret. Iss. Ergon. Sci.* 22, 488–509. doi: 10.1080/1463922X.2020.1827080

Buchanan, M. (2017). Colonizing mars. *Nat. Phys.* 13, 1035–1035. doi: 10.1038/nphys4311

Calhoun, G. L. (2022). Adaptable (not adaptive) automation: forefront of human-automation teaming. *Hum. Fact.* 64, 269–277. doi: 10.1177/00187208211037457

Carter-Browne, B. M., Paletz, S. B., Campbell, S. G., Carraway, M. J., Vahlkamp, S. H., Schwartz, J., et al. (2021). *There is No "AI" in Teams: A Multidisciplinary Framework for AIs to Work in Human Teams*. Maryland, MA: Applied Research Laboratory for Intelligence and Security (ARLIS).

Chen, M., Nikolaidis, S., Soh, H., Hsu, D., and Srinivasa, S. (2018). Planning with trust for human-robot collaboration. *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (Chicago, IL), 307–315. doi: 10.1145/3171221.3171264

Cordes, F., Kirchner, F., and Bremen, R. I. C. (2010). Heterogeneous robotic teams for exploration of steep crater environments. *International Conference on Robotics and Automation (ICRA)*, Anchorage, Alaska.

Demiris, Y. (2007). Prediction of intent in robotics and multi-agent systems. *Cogn. Process.* 8, 151–158. doi: 10.1007/s10339-007-0168-9

Dettmann, A., Voegele, T., Ocón, J., Dragomir, I., Govindaraj, S., De Benedetti, M., et al. (2022). *COROB-X: A Cooperative Robot Team for the Exploration of Lunar Skylights*. Available online at: https://hdl.handle.net/10630/24387 (accessed September 11, 2023).

DLR (2020). *Auch CIMON-2 Meistert Seinen Einstand auf der ISS*. Available online at: https://www.dlr.de/de/aktuelles/nachrichten/2020/02/20200415_auch-cimon-2-meistert-seinen-einstand-auf-der-iss (accessed September 11, 2023).

Endsley, M. R. (1999). "Situation awareness in aviation systems," in *Handbook of Aviation Human Factors*, eds D. J. Garland, J. A. Wise, and V. D. Hopkin (Mahwah, NJ: Lawrence Erlbaum Associates Publishers), 257-276.

Flin, R., O'Connor, P., and Crichton, M. (2008). *Safety at the Sharp End*. London: Ashgate.

Govindarajan, V., Bhattacharya, S., and Kumar, V. (2016). "Human-robot collaborative topological exploration for search and rescue applications," in Cho *Distributed Autonomous Robotic Systems*, eds N. Y. Chong and Cho Y.-J. (Springer), 17-32. doi: 10.1007/978-4-431-55879-8_2

Gutzeit, L., and Kirchner, F. (2022). Unsupervised segmentation of human manipulation movements into building blocks. *IEEE Access* 10, 125723–125734. doi: 10.1109/ACCESS.2022.3225914

Hackman, J. R. (1987). "The design of work teams," in *Handbook of Organizational Behavior*, ed J.W. Lorsch (Englewood Cliffs, NJ: Prentice-Hall), 315–342.

Hagemann, V., and Kluge, A. (2017). Complex problem solving in teams: the influence of collective orientation on team process demands. *Front. Psychol. Cog. Sci.* 8, 1730. doi: 10.3389/fpsyg.2017.01730

Hagemann, V., Ontrup, G., and Kluge, A. (2021). Collective orientation and its implications for coordination and team performance in interdependent work contexts. *Team Perform. Manag.* 27, 30–65. doi: 10.1108/TPM-03-2020-0020

Hambuchen, K., Marquez, J., and Fong, T. (2021). A review of NASA human-robot interaction in space. *Curr. Robot. Rep.* 2, 265–272. doi: 10.1007/s43154-021-00062-5

Hoff, K. A., and Bashir, M. (2015). Trust in automation: integrating empirical evidence on factors that influence trust. *Hum. Fact.* 57, 407–434. doi: 10.1177/0018720814547570

IBM and ver.di (2020). Künstliche Intelligenz: Ein sozialpartnerschaftliches Forschungsprojekt untersucht die neue Arbeitswelt. Available online at: https://www.verdi.de/$++$file$++$5fc901bc4ea3118def3edd33/download/20201203_KI-Forschungsprojekt-verdi-IBM-final.pdf (accessed September 11, 2023).

Kabir, A. M., Kanyuck, A., Malhan, R. K., Shembekar, A. V., Thakar, S., Shah, B. C., et al. (2019). Generation of synchronized configuration space trajectories of multi-robot systems. *International Conference on Robotics and Automation (ICRA)* (Montreal: IEEE), 8683-8690. doi: 10.1109/ICRA.2019.8794275

Kirchner, F. (2020). AI-perspectives: the turing option. *AI Perspect.* 2, 2. doi: 10.1186/s42467-020-00006-3

Kluge, A., Hagemann, V., and Ritzmann, S. (2014). "Military crew resource management – Das Streben nach der bestmöglichen Teamarbeit [Striving for the best of teamwork]," in *Psychologie für Einsatz und Notfall [Psychology for mission and emergency]*, eds G. Kreim, S. Bruns, and B. Völker (Bernard and Graefe in der Mönch Verlagsgesellschaft mbH), 141–152.

Lyons, J. B., Jessup, S. A., and Voc, T. Q. (2022). The role of decision authority and stated social intent as predictors of trust in autonomous robots, *Top. Cogn. Sci.* 4, 1–20. doi: 10.1111/tops.12601

Marks, M. A., Mathieu, J. E., and Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. *Acad. Manag. Rev.* 26, 356–376. doi: 10.2307/259182

Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Salas, E., and Cannon-Bowers, J. A. (2000). The influence of shared mental models on team process and performance. *J. Appl. Psychol.* 85, 273–283. doi: 10.1037/0021-9010.85.2.273

McNeese, N. J., Demir, M., Chiou, E. K., and Cooke, N. J. (2021a). Trust and team performance in human–autonomy teaming. *Int. J. Elect. Comm.* 25, 51–72. doi: 10.1080/10864415.2021.1846854

McNeese, N. J., Demir, M., Cooke, N. J., and She, M. (2021b). Team situation awareness and conflict: a study of human–machine teaming. *J. Cogn. Engin. Dec. Mak.* 15, 83–96. doi: 10.1177/15553434211017354

NASA (2017). *NASA's Journey to Mars*. Available online at: https://www.nasa.gov/content/nasas-journey-to-mars (accessed September 11, 2023).

National Academies of Sciences, Engineering, and Medicine (2022). *Human-AI Teaming: State-of-the-Art and Research Needs*. Washington, DC: The National Academies Press.

Niewiadomski, R., Suresh, A., Sciutti, A., and Di Cesare, G. (2021). Vitality forms analysis and automatic recognition. *J. Latex Class Files* 14, 1. doi: 10.36227/techrxiv.16691476.v1

O'Neill, T. A., Flathmann, C., McNeese, N. J., and Salas, E. (2023). Human-autonomy teaming: need for a guiding team-based framework? *Comput. Human Behav.* 146, 107762. doi: 10.1016/j.chb.2023.107762

O'Neill, T. A., McNeese, N., Barron, A., and Schelble, B. (2022). Human–autonomy teaming: a review and analysis of the empirical literature. *Hum. Fact.* 64, 904–938. doi: 10.1177/0018720820960865

Paoletti, J., Kilcullen, M., and Salas, E. (2021). "Teamwork in space exploration," in *Psychology and Human Performance in Space Programs*, eds L. B. Landon, K. Slack, and E. Salas (CRC Press), 195-216. doi: 10.1201/9780429440878-10

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., Bernstein, M. S., et al. (2023). Generative agents: interactive simulacra of human behavior. *arXiv preprint* 2304, 03442. doi: 10.48550/arXiv.2304.03442

Raileanu, R., Denton, E., Szlam, A., and Fergus, R. (2018). Modeling others using oneself in multi-agent reinforcement learning. *International Conference on Machine Learning* (Stockholm: PMLR), 4257-4266.

Rhinehart, N., McAllister, R., Kitani, K., and Levine, S. (2019). "Precog: prediction conditioned on goals in visual multi-agent settings," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2821-2830. doi: 10.1109/ICCV.2019.00291

Rieth, M., and Hagemann, V. (2022). Automation as an equal team player for humans? A view into the field and implications for research and practice. *Appl. Ergon.* 98, 103552. doi: 10.1016/j.apergo.2021.103552

Roehr, T. M., Cordes, F., and Kirchner, F. (2014). Reconfigurable integrated multirobot exploration system (RIMRES): heterogeneous modular reconfigurable robots for space exploration. *J. Field Robot.* 31, 3–34. doi: 10.1002/rob.21477

Salas, E., Sims, D., and Burke, S. (2005). Is there a "Big Five" in Teamwork? *Small Group Res.* 36, 555–599. doi: 10.1177/1046496405277134

Salas, E., Tannenbaum, S. I., Kozlowski, S. W. J., Miller, C. A., Mathieu, J. E., Vessey, W. B., et al. (2015). Teams in space exploration: a new frontier for the science of team effectiveness. *Curr. Dir. Psychol. Sci.* 24, 200–207. doi: 10.1177/0963721414566448

Schecter, A., Hohenstein, J., Larson, L., Harris, A., Hou, T-. Y., Lee, W-. Y., et al. (2022). Vero: an accessible method for studying human-AI teamwork. *Comput. Human Behav.* 141, 107606. doi: 10.1016/j.chb.2022.107606

Schneider, M., Miller, M., Jacques, D., Peterson, G., and Ford, T. (2021). Exploring the impact of coordination in human–agent teams. *J. Cogn. Engin. Dec. Mak.* 15, 97–115. doi: 10.1177/15553434211010573

Schraagen, J. M., Barnhoorn, J. S., van Schendel, J., and van Vught, W. (2022). Supporting teamwork in hybrid multi-team systems. *Theoret. Iss. Ergon. Sci.* 23, 199–220. doi: 10.1080/1463922X.2021.1936277

Schwartz, T., Feld, M., Bürckert, C., Dimitrov, S., Folz, J., Hutter, D., et al. (2016a). Hybrid Teams of humans, robots and virtual agents in a production setting. *Proceedings of the 12th International Conference on Intelligent Environments, (IE-16), 12, 9.-13.9.2016* (London: IEEE). doi: 10.1109/IE.2016.53

Schwartz, T., Zinnikus, I., Krieger, H. U., Bürckert, C., Folz, J., Kiefer, B., et al. (2016b). "Hybrid teams: flexible collaboration between humans, robots and virtual agents," in *Proceedings of the 14th German Conference on Multiagent System Technologies*, eds M. Klusch, R. Unland, O. Shehory, A. Pokhar, and S. Ahrndt (Klagenfurt: Springer, Series Lecture Notes in Artificial Intelligence), 131-146.

Seeber, I., Bittner, E., Briggs, R. de Vreede, T., de Vreede, G.-J., Elkins, G. J., et al. (2020). Machines as teammates: a research agenda on AI in team collaboration. *Inform. Manag.* 57, 1–22. doi: 10.1016/j.im.2019.103174

Smith-Jentsch, K. A., Baker, D. P., Salas, E., and Cannon-Bowers, J. A. (2001). "Uncovering differences in team competency requirements: the case of air traffic control teams.," in *Improving Teamwork in Organizations. Applications of Resource Management Training*, eds E. Salas, C. A. Bowers, and E. Edens (Mahwah, NJ: Lawrence Erlbaum Associates Publishers), 31-54.

Stowers, K., Brady, L. L., MacLellan, C., Wohleber, R., and Salas, E. (2021). Improving teamwork competencies in human-machine teams: perspectives from team science. *Front. Psychol.* 12, 590290. doi: 10.3389/fpsyg.2021.590290

Tabrez, A., Luebbers, M. B., and Hayes, B. (2020). A survey of mental modeling techniques in human–robot teaming. *Curr. Rob. Rep.* 1, 259–267. doi: 10.1007/s43154-020-00019-0

Taylor, A. (2021). *Human-Robot Teaming in Safety-Critical Environments: Perception of and Interaction with Groups* (Publication No. 28544730) [Doctoral dissertation, University of California]. ProQuest Dissertations and Theses Global.

Teahan, W. J. (2010). *Artificial Intelligence—Agent Behaviour*. Bookboon.

Trick, S., Koert, D., Peters, J., and Rothkopf, C. A. (2019). Multimodal uncertainty reduction for intention recognition in human-robot interaction. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Macau), 7009-7016. doi: 10.1109/IROS40897.2019.8968171

Turing, A. (1948). *Intelligent Machinery*. New York, NY: B. Jack Copeland.

Uitdewilligen, S., and Waller, M. J. (2018). Information sharing and decision making in multidisciplinary crisis management teams. *J. Organ. Behav.* 39, 731–748. doi: 10.1002/job.2301

Ulfert, A-. S., Antoni, C. H., and Ellwart, T. (2022). The role of agent autonomy in using decision support systems at work. *Comput. Human Behav.* 126, 106987. doi: 10.1016/j.chb.2021.106987

Vezhnevets, A. S., Osindero, S., Schaul, T., Heess, N., Jaderberg, M., Silver, D., et al. (2017). Feudal networks for hierarchical reinforcement learning. *International Conference on Machine Learning* (Sydney), 3540–3549.

Waller, M. J., Gupta, N., and Giambatista, R. C. (2004). Effects of adaptive behaviors and shared mental models on control crew performance. *Manage. Sci.* 50, 1534–1544. doi: 10.1287/mnsc.1040.0210

Wang, L., Gao, R., Váncza, J., Krüger, J., Wang, X. V., Makris, S., et al. (2019). Symbiotic human-robot collaborative assembly. *CIRP Annals* 68, 701–726. doi: 10.1016/j.cirp.2019.05.002

Westbrook, L. (2006). Mental models: a theoretical overview and preliminary study. *J. Inform. Sci.* 32, 563–579. doi: 10.1177/0165551506068134

Xia, Z., Lei, Q., Yang, Y., Zhang, H., He, Y., Wang, W., et al. (2019). Vision-based hand gesture recognition for human-robot collaboration: a survey. *5th International Conference on Control, Automation and Robotics (ICCAR),* (Beijing) 198-205. doi: 10.1109/ICCAR.2019.8813509

Zaccaro, S. J., Marks, M. A., and DeChurch, L. A. (2012). "Multiteam systems: an introduction," in *Multiteam Systems: An Organization Form for Dynamic and Complex Environments*, eds S. J. Zaccaro, M. A. Marks, and L. A. DeChurch (New York, NY: Routledge), 3–32. doi: 10.4324/9780203814772

frontiers | Frontiers in Artificial Intelligence

TYPE Systematic Review
PUBLISHED 29 September 2023
DOI 10.3389/frai.2023.1250725

OPEN ACCESS

EDITED BY
Margaret A. Goralski,
Quinnipiac University, United States

REVIEWED BY
Krystyna Gorniak-Kocikowska,
Southern Connecticut State University, United States
Pranav Gupta,
University of Illinois at Urbana–Champaign, United States

*CORRESPONDENCE
Sophie Berretta
✉ sophie.berretta@rub.de
Alina Tausch
✉ alina.tausch@rub.de

†These authors have contributed equally to this work and share first authorship

RECEIVED 30 June 2023
ACCEPTED 11 September 2023
PUBLISHED 29 September 2023

CITATION
Berretta S, Tausch A, Ontrup G, Gilles B, Peifer C and Kluge A (2023) Defining human-AI teaming the human-centered way: a scoping review and network analysis. *Front. Artif. Intell.* 6:1250725. doi: 10.3389/frai.2023.1250725

Defining human-AI teaming the human-centered way: a scoping review and network analysis

Sophie Berretta[1*†], Alina Tausch[1*†], Greta Ontrup[1], Björn Gilles[1], Corinna Peifer[2] and Annette Kluge[1]

[1]Department of Psychology, Organizational, and Business Psychology, Ruhr University Bochum, Bochum, Germany, [2]Department of Psychology I, University of Lübeck, Lübeck, Germany

Introduction: With the advancement of technology and the increasing utilization of AI, the nature of human work is evolving, requiring individuals to collaborate not only with other humans but also with AI technologies to accomplish complex goals. This requires a shift in perspective from technology-driven questions to a human-centered research and design agenda putting people and evolving teams in the center of attention. A socio-technical approach is needed to view AI as more than just a technological tool, but as a team member, leading to the emergence of human-AI teaming (HAIT). In this new form of work, humans and AI synergistically combine their respective capabilities to accomplish shared goals.

Methods: The aim of our work is to uncover current research streams on HAIT and derive a unified understanding of the construct through a bibliometric network analysis, a scoping review and synthetization of a definition from a socio-technical point of view. In addition, antecedents and outcomes examined in the literature are extracted to guide future research in this field.

Results: Through network analysis, five clusters with different research focuses on HAIT were identified. These clusters revolve around (1) human and (2) task-dependent variables, (3) AI explainability, (4) AI-driven robotic systems, and (5) the effects of AI performance on human perception. Despite these diverse research focuses, the current body of literature is predominantly driven by a technology-centric and engineering perspective, with no consistent definition or terminology of HAIT emerging to date.

Discussion: We propose a unifying definition combining a human-centered and team-oriented perspective as well as summarize what is still needed in future research regarding HAIT. Thus, this work contributes to support the idea of the Frontiers Research Topic of a theoretical and conceptual basis for human work with AI systems.

KEYWORDS

artificial intelligence, human-centered AI, network analysis, bibliometric analysis, bibliometric coupling, work psychology, human-AI teaming, humane work

1. Introduction

With the uprise of technologies based on artificial intelligence (AI) in everyday professional life (McNeese et al., 2021), human work is increasingly affected by the use of AI, with the growing need to cooperate or even team up with it. AI technologies describe intelligent systems executing human cognitive functions such as learning, interacting,

solving problems, and making decisions, which is an enabler for using them in a similarly flexible manner as human employees (e.g., Huang et al., 2019; Dellermann et al., 2021). Thus, the emerging capabilities of AI technologies allow them to be implemented directly in team processes with other artificial and human agents or to overtake functions that support humans in a way team partners would. Such can be referred to as human-AI teaming (HAIT; McNeese et al., 2018). HAIT constitutes a human-centered approach to AI implementation at work, as its aspiration is to leverage the respective strengths of each party. The diverse but complementary capabilities of human-AI teams foster effective collaboration and enable the achievement of complex goals while ensuring human wellbeing, motivation, and productivity (Kluge et al., 2021). Other synergies resulting from human-AI teaming facilitate strategic decision making (Aversa et al., 2018), the development of individual capabilities, and thus employee motivation in the long term (Hughes et al., 2019).

Up to now, the concept of HAIT has been investigated from various disciplinary perspectives, e.g., engineering, data sciences or psychology (Wilkens et al., 2021). An integration of these perspectives seems necessary at this point to design complex work systems as human-AI teams with technical, human, task, organizational, process-related, and ethical factors in mind (Kusters et al., 2020). In addition to this, a conceptual approach with a unifying definition is needed to unite research happening under different terms, but with a potentially similar concept behind it. To evolve from multi- to interdisciplinarity, the field of HAIT research needs to overcome several obstacles:

(1) The discipline-specific definitions and understandings of HAIT have to be brought together or separated clearly.
(2) Different terms used for the same concept, e.g., human-autonomy teaming (O'Neill et al., 2022) and human-AI collaboration (Vössing et al., 2022), have to be identified to enable knowledge transfer and integration of empirical and theoretical work.
(3) The perspectives on either the technology or the human should be seen as complementary, not as opposing.

As "construct confusion can [...] create difficulty in building a cohesive body of scientific literature" (O'Neill et al., 2022, p. 905), it is essential that different disciplines find the same language to talk about the challenges of designing, implementing and using AI as a teammate at work. Therefore, the goal of this scoping review is to examine the extent, range, and nature of current research activities on HAIT. Specifically, we want to give an overview of the definitory understandings of HAIT and of the current state of empirically investigated and theoretically discussed antecedents and outcomes within the different disciplines. Based on a bibliometric network analysis, research communities will be mapped and analyzed regarding their similarities and differences in the understanding of HAIT and related research activities. By this, our scoping review reaches synergistic insights and identifies research gaps in examining human-AI teams, promoting the formation of a common understanding.

2. Theoretical background: human-AI teaming in the workplace

As technologies progress and AI becomes more widely applied, humans will no longer work together only with other humans but will increasingly need to use, interact with and leverage AI technologies to achieve complex goals. Increasingly "smart" AI technologies entail characteristics that require new forms of work and cooperation between human and technology (Wang et al., 2021), developing from "just" technological tools to teammates to human workers (Seeber et al., 2020). According to the CASA-paradigm, people tend to perceive *c*omputers *a*s *s*ocial *a*ctors (Nass et al., 1996), which is probably even more true with highly autonomous technologies driven by AI, being seen as very agentic. This opens opportunities to move the understanding of AI as a helpful technological application to a team member that interdependently works with the employee toward a shared and valued goal (Rix, 2022). Thus, human-AI teams evolve as a new form of work, pairing human workforce and abilities with that of AI.

Why is a shift in parameters needed? Our proposed answer is that that it offers a new, humane attempt toward AI implementation at work that respects employees' needs, feeling of belongingness and experience (Kluge et al., 2021). Additionally, employees' acceptance, and a positive attitude in working with an AI can improve when it is seen as a teammate (see, e.g., Walliser et al., 2019). Thus, HAIT provides an opportunity to create attractive and sustainable workplaces by harnessing people's capabilities and enabling learning and mutual support. This in turn leads to synergies (Kluge et al., 2021), increased motivation and wellbeing on the part of humans, by spending more time on identity-forming and creative tasks, while safety-critical and monotonous tasks can be handed over to the technology (Jarrahi, 2018; Kluge et al., 2021; Berretta et al., 2023). In addition to the possibility of creating human-centered workplaces, the expected increase in efficiency and performance due to complementary capabilities of humans and AI technologies, described as synergies, are further important reasons for the parameter shift (Dubey et al., 2020; Kluge et al., 2021).

However, those advantages connected to the human workforce and the performance do not just come naturally when pairing humans with AI systems. The National Academies of Sciences, Engineering, and Medicine (2021) defines four conditions for a human-AI team to profit from these synergies:

(1) The human part has to be able to understand and anticipate the behaviors of the deployed intelligent agents.
(2) To ensure appropriate use of AI systems, the human should be able to establish an appropriate relationship of trust.
(3) The human part can make accurate decisions when using the output information of the deployed systems and
(4) has the ability to control and handle the systems appropriately.

These conditions demonstrate that successful teaming depends on technical (e.g., design of the AI system) as well as human-related dimensions (e.g., trust in the system) and additionally requires interaction/teamwork issues (e.g., form of collaboration). This makes HAIT an inherently multidisciplinary field, that

should be explored in the spirit of joint optimization to achieve positive results in all dimensions (Vecchio and Appelbaum, 1995). Nevertheless, joint consideration and optimization is still not common practice in the development of technologies or the design of work systems (Parker et al., 2017), so that much research looks at HAIT solely from one perspective. The following section introduces two perspectives on teams in work contexts relevant for the proposed, joint HAIT approach.

2.1. Human-technology teaming

The field of human-technology teaming encompasses a number of established concepts, including human-machine interaction (e.g., Navarro et al., 2018) or human-automation interaction (e.g., Parasuraman et al., 2000). These constructs can, but do not have to, include aspects of teaming: they describe a meta-level of people working in some kind of contact with technologies. Concepts further specify on two different aspects: the interaction aspect and the technology aspect. The term "interaction" as a broad concept is increasingly replaced by terms trying to detail the type of interaction such as co-existence, cooperation and collaboration (Schmidtler et al., 2015), usually understood as increasingly close and interdependent contact. Maximally interdependent collaboration including an additional aspect of social bonding (team or group cohesion, see Casey-Campbell and Martens, 2009) is called teaming. In terms of the technology aspect, a range of categories exists from general terms like technology, machines or automation, which can be broad or specific, depending on the context (Lee and See, 2004). More specific categories include autonomy, referring to adaptive, self-governed learning technologies (Lyons et al., 2021), robots or AI.

A recent and central concept in this research field is human-autonomy teaming, as introduced by O'Neill et al. (2022) in their review. Although using a different term than HAIT, this concept plays a crucial role in consolidating and unifying research on the teaming of humans and autonomous, AI-driven systems. Their defining elements of human-autonomy teaming include:

(1) a machine with high agency,
(2) communicativeness of the autonomy,
(3) conveying information about its intent,
(4) evolving shared mental models,
(5) and interdependence between humans and the machines (O'Neill et al., 2022).

However, there are several critical aspects to consider in this review: The term "human-autonomy teaming" can elicit associations that may not contribute to the construct of HAIT. The definition of autonomy varies between different fields and the term alone can be misleading, as it can be understood as the human's autonomy, the autonomy of a technical agent, or as the degree of autonomy in the relationship. Additionally, O'Neill et al.'s (2022) reliance on the levels of automation concept (Parasuraman et al., 2000) reveals a blind spot in human-centeredness, because the theory fails to consider different perspectives (Navarro et al., 2018) and is not selective enough to describe complex human-machine interactions. Furthermore, the review primarily focuses on empirical research, neglecting conceptual work on teaming between humans and autonomous agents. As a result, the idea of teaming is—despite the name—not as prominent as expected, and the dynamic, mutually supportive aspect of teams is overshadowed by the emphasis on technological capabilities for human-autonomy teaming.

In addition to the emerging problem of research focusing solely on technology aspects, which is important, but insufficient to fully describe and understand a multidimensional system like HAIT, different definitions exist to describe what we understand by human-AI teams. Besides the already mentioned definition of human-autonomy teaming, Cuevas et al. (2007) for example describe HAIT as "one or more people and one or more AI systems requiring collaboration and coordination to achieve successful task completion" (p. 64). Demir et al. (2021, p. 696) define that in HAIT "human and autonomous teammates promptly interact with one another in response to information flow from one team member to another, adapt to the dynamic task, and achieve common goals". While these definitions share elements, such as the idea of working toward a common goal with human and autonomous agents, there are also dissimilarities among the definitions, for example, in the terminology used, as seemingly similar terms like interaction and collaboration represent different constructs (Wang et al., 2021).

In an evolving research field, terminology ambiguity can inspire different research foci, but also pose challenges. Different emerging research fields might refer to the same phenomenon using various terms (i.e., human-AI-teaming vs. human-autonomy-teaming or interaction vs. teaming), which is known as jangle-fallacy and can cause problems in research (Flake and Fried, 2020). Such conceptual blurring may hinder interdisciplinary exchange and the integration of findings from different disciplines due to divergent terminology (O'Neill et al., 2022).

2.2. Human-human teaming

Another important perspective to consider is that of human teams, which forms the foundation of team research. Due to its roots in psychology and social sciences, the perspective on teams is traditionally a human-centered one, implying relevant insights on the blind spot of human-technology teaming research. The term "team" refers to two or more individuals interacting independently to reach a common goal and experiencing a sense of "us" (Kauffeld, 2001). Each team member is assigned a specific role or function, usually for a limited lifespan (Salas et al., 2000). Teamwork allows for the combination of knowledge, skills, and specializations, the sharing of larger tasks, mutual support in problem-solving or task execution, and the development of social structures (Kozlowski and Bell, 2012).

The roots of research on human teams can be traced to the Hawthorne studies conducted in the 1920s and 1930s (Mathieu et al., 2017). Originally designed to examine the influence of physical work conditions (Roethlisberger and Dickson, 1939), these studies unexpectedly revealed the impact of group dynamics on performance outcomes, leading to a shift in focus toward interpersonal relationships between workers and managers (Sundstrom et al., 2000). In this way, psychology's understanding

of teamwork and its effects has since stimulated extensive theory and research on group phenomena in the workplace (Mathieu et al., 2017). Following over a century of research, human teamwork, once a "black box" (Salas et al., 2000, p. 341), is now well-defined and understood. According to Salas et al. (2000), teams are characterized by three main elements: Firstly, team members have to be able to coordinate and adapt to each other's requirements in order to work effectively as a team. Secondly, communication between team members is crucial, particular in uncertain and dynamic environments, where information exchange is vital. Lastly, a shared mental model is essential for teamwork, enabling team members to align their efforts toward a common goal and motivate each other. Moreover, successful teamwork requires specific skills, such as adaptability, shared situational awareness, team management, communication, decision-making, coordination, feedback, and interpersonal skills (Cannon-Bowers et al., 1995, see Supplementary Table 1 for concept definitions).

Commonalities of human-human teams and human-AI teams have already been identified in terms of relevant features and characteristics that contribute to satisfactory performance, including shared mental models, team cognitions, situational awareness and communication (Demir et al., 2021). Using human-human teams research insights as a basis for HAIT offers access to well-established and tested theories and definitions, but leaves unclarities in the questions which characteristics and findings can be effectively transferred to HAIT research and what the vital existing differences are (McNeese et al., 2021).

2.3. Combining human-technology and human-human teaming in a human-centered way

A consideration of both the human-human and human-technology teaming perspectives serves as a useful and necessary starting point for exploring human-AI teams. In order to advance our understanding, it is crucial to combine the findings from these perspectives and integrate them within a socio-technical systems approach. The concept of socio-technical systems recognizes that the human part is intricately linked to the technological elements in the workplace, with both systems influencing and conditioning each other (Emery, 1993). Therefore, a comprehensive understanding of human-AI teams can only be achieved through an integrative perspective that considers the interplay between humans and technology, as well as previous insights from both domains regarding teaming. In our review, we aim to address the lack of integration by…

- establishing the term human-AI teaming (HAIT) as an umbrella term for teamwork with any sort of artificially intelligent (partially), autonomously acting system.
- omitting a theoretical basement for embedding our literature search and analysis. We want to neutrally identify how (different) communities understand and use HAIT and what might be the core to it, without pre-assumptions on the characteristics.
- taking a human-centered perspective and using the ideas of socio-technical system designs to discuss our findings, anyways.
- including a broad range of scientific literature, which contains conceptual and theoretical papers—thereby being able to cover a deeper examination of HAIT-related constructs.
- seeing if the understanding of teaming has developed since the review by O'Neill et al. (2022) and if there are papers considering especially the team level and dynamics associated with agents sharing tasks.

2.4. Rational for this study: research questions and intentions

The goal of this paper is to examine the scope, breadth, and nature of the most current research on HAIT. In this context, we are interested in understanding the emerging research field, the streams and disciplines involved, by visualizing and analyzing current research streams using clusters based on a bibliometric network analysis ("who cites who"). The aim is to use mathematical methods to capture and analyze the relationships between pieces of literature, thereby representing the quantity of original research and its citation dependencies to related publications (Kho and Brouwers, 2012). The investigation of resulting networks can reveal research streams and trends in terms of content and methodology (Donthu et al., 2020). Concisely, the objective of the network analysis is to investigate the following research question:

RQ1: Which clusters can be differentiated regarding interdisciplinary and current human-AI teaming research based on their relation in the bibliometric citation network?

Further, the publications of the identified clusters will be examined based on a scoping review concerning the definitory understanding of human-AI teams as well as their empirically investigated or theoretically discussed antecedents and outcomes. This should contribute to answering the subsequent research questions:

RQ2: Which understandings of human-AI teaming emerge from each cluster in the network?

RQ3: Which antecedents and outcomes of human-AI teaming are currently empirically investigated or theoretically discussed?

This second part of the analysis should lead to a consideration of the quality of publications in the network in addition to the quantity within the network analysis (Kho and Brouwers, 2012). We want to give an overview of what is seen as the current core of HAIT within different research streams and identify differences and commonalities. On the one hand, making differences in the understanding of HAIT explicit is important, as it allows future research to develop into decidedly distinct research strands. On the other hand, the identification of similarities creates a basis for the development of a common language about HAIT, which will allow the establishment of common ground in the future so that the interdisciplinary exchange on what HAIT is and can be grows in stringency. To also contribute to this aspect, we aim to identify a definition of HAIT that serves the need for a common ground. In doing so, the definition is intended to extend that of O'Neill et al. (2022), reflecting the latest state of closely related research as well

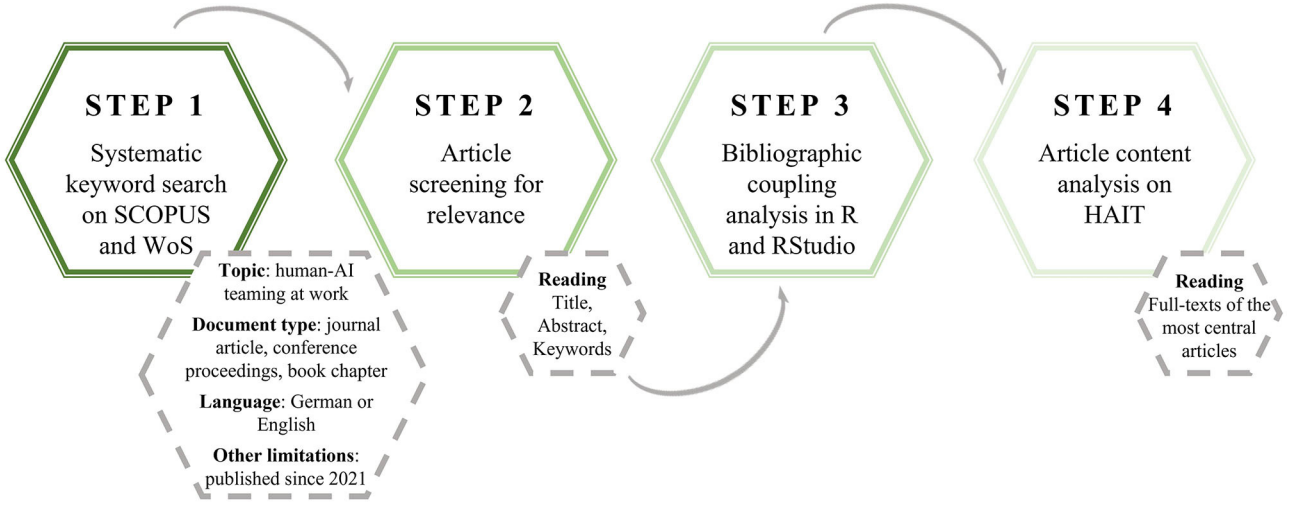

FIGURE 1
Illustration of the procedure. WoS, web of science; HAIT, human-AI teaming.

as addressing and considering the problems identified earlier. If we are not able to find this kind of a definition within the literature that focuses on the teaming aspect, we want to use the insights from our research to newly develop such a definition of HAIT. Thus, our fourth research question, which we will be able to answer after collecting all other results and discussing their implications, is:

RQ4: How can we define HAIT in a way that is able to bridge different research streams?

This is expected to help researchers from different disciplines finding a shared ground in definitions and concepts and explicating divergences in understanding. By identifying the current state of research streams and corresponding understandings of HAIT, as well as the antecedents and outcomes, synergistic insights and research gaps can be identified. A unifying definition will further help stimulate and align further research on this topic.

3. Materials and methods

To identify research networks and to analyze their findings on HAIT, the methods of bibliometric network analysis and scoping review were combined. The pre-registration for this study can be accessed here: https://doi.org/10.23668/psycharchives.12496.

3.1. Literature search

The basis for the network analysis and the scoping review was a literature search in Clarivate Analytic's Web of Science (WoS) and Elsevier's Scopus (Scopus) databases. Those were chosen because they represent the main databases for general-purpose scientific publications, spanning articles, conference proceedings and more (Kumpulainen and Seppänen, 2022). The process of the literature search was conducted and is reported according to the PRISMA reporting Guidelines for systematic reviews (Moher et al., 2009), more specifically the extension for scoping reviews (PRISMA-ScR; Tricco et al., 2018). Figure 1 provides an overview of the integrated procedure.

The literature search was conducted on the 25.01.2023. The keywords for our literature search (see Table 1) were chosen to include all literature in the databases that relates to HAIT in the workplace. Thus, the components "human," "AI," "teamwork," and "work" all needed to be present in any (synonymous) form. Furthermore, only articles published since the year 2021 were extracted. This limited time frame was chosen as the goal was

TABLE 1 Used search-terms for the database-search.

Human	AI	Teaming	Work
"human"	"AI"	"team*"	"work"
"people"	"artificial intelligence"	"collaborat*"	"occupation"
"employee"	"machine learning"	"cooperation"	"profession"
"Mensch"	"synthetic agent"	"symbiosis"	"job"
"Mitarbeiter*"	"autonomous agent"	"alliance"	"Arbeit"
"Beschäftigte*"	"KI"	"coalition"	"Beruf*"
	"Künstliche Intelligenz"	"partner*"	
	"maschinelles lernen"	"Kollaboration"	
		"Kooperation"	
		"Symbiose"	
		"Tandem"	

Four categories of terms were used: Human, AI, Teaming and Work. Terms inside a category were connected by the search operator "OR" and Categories themselves were connected by the "AND" operator. The asterisk serves as a wildcard for different endings to a common word stem.

to map the most current research front, using the European industrial strategy "Industry 5.0" (Breque et al., 2021) as s starting point. Its focus on humans, their needs, and capabilities instead of technological system specifications represents a shift in attention to the individual that is accompanied by the explicit mention of creating a team of human(s) and technical system(s) (Breque et al., 2021), therefore marking a good starting point of a joint human-AI teaming understanding. Accordingly, only the most current literature published since the introduction of Industry 5.0 and not yet included in the review of O'Neill and colleagues is taken into account in our review (note that by analyzing the references in bibliometric coupling and qualitatively evaluating the referred concepts of HAIT, we also gain information on older important literature). Included text types were peer-reviewed journal articles, conference proceedings and book chapters (not limited to empirical articles) in English or German language. As shown in the PRISMA-diagram in Figure 2, the search resulted in $n = 1{,}963$ articles being retrieved. After removing $n = 440$ duplicates, abstract-screening was conducted using the web-tool Rayyan (Ouzzani et al., 2016). In case of duplicates, the WoS version was kept for its preferable data structure.

Six researchers familiar with the subject screened the abstracts, with every article being judged by at least two blind raters. Articles not dealing with the topic "human-AI teaming in the workplace" or being incorrectly labeled in the database and not fitting our eligibility criteria were excluded. In case of disagreement or uncertainty, raters discussed and compared their reasoning and decided on a shared decision, and/or consulted the other raters. In total, 1,159 articles (76%) were marked for exclusion. Exclusion criteria were: (a) publication in another language than English or German, (b) publication form of a book (monography or anthology), (c) work published before 2021, (d) work not addressing human-AI-teaming in title, abstract or keywords, (e) work not addressing work context in title, abstract or keywords. The remaining articles ($n = 364$, 24%) were used in the network analysis (see Figure 2).

3.2. Bibliometric mapping approach and clustering algorithm

To map and cluster the included literature and thus describe the network that structures the research field of HAIT, a bibliometric mapping approach and clustering algorithm had to be chosen. Networks consist of publications that are mapped, called *nodes*, and the connection between those nodes, which are called *edges* (Hevey, 2018). Which publications appear in the nodes and how the edges are formed depends on the mapping approach used. The variety includes direct citation, bibliometric coupling, and co-citation networks (Boyack and Klavans, 2010), but bibliometric coupling analysis has been shown to be the most accurate (Boyack and Klavans, 2010). It works by first choosing a sample of papers, serving as the network nodes. The edges are then created by comparing the references of the node-papers, adding edges between two publications if they share references (Jarneving, 2005). Thus, the newest publications are mapped, while the cited older publications themselves are not included in the network (Boyack and Klavans, 2010; Donthu et al., 2021). Since our goal was to map and cluster the current research front, we chose bibliometric coupling for our network analysis approach.

The article metadata from WoS and Scopus were prepared for network analysis using R (version 4.2, R Core Team, 2022), as well as their reference lists. We did this in a way that the first author, including initials, the publishing year, the starting page, and the volume were extracted from all cited references. This information was then combined in a new format string. In total, $n = 17.323$ references containing at least first author and year were generated. Of those, 8.955 references contained missing data about the starting page, the volume or both. To minimize the risk of two different articles randomly having the same reference string, we excluded all references that missed both volume and starting page information ($n = 3{,}794$). We kept all references that only had either starting page ($n = 1{,}384$) or volume ($n = 3{,}794$) information missing, due to a low probability and influence of single duplicates.

3.3. Network analysis

Using the newly created format, we conducted a coupling network analysis using R and the packages *igraph* (version 1.3.1; Aria and Cuccurullo, 2017) and *bibliometrix* (version 4.1.0; Csárdi et al., 2023). The used code can be accessed here: https://github.com/BjoernGilles/HAIT-Network-Analysis. Bibliometrix was used to create the first weighted network with no normalization. Then it was converted into igraph format, removing any isolated edges with degree = 0. The degree centrality refers to the number of edges a node is connected by to other nodes, while the weighted degree centrality adapts this measure by multiplying it with the strength of the edge (Donthu et al., 2021). Then, the multilevel community clustering algorithm was used to identify the dominant clusters. Multilevel-clustering was chosen since the network's mixing parameter was impossible to predict a-priori and since it shows stable performance for a large range of clustering structures (Yang et al., 2016). The stability of our clustering solution was checked by comparing our results with 10,000 recalculations of the multilevel-algorithm on our network-data.

Afterwards, all clusters containing ≥ 20 nodes were selected and split into subgraphs. The top 10% of papers with the highest weighted degree of each subgraph were selected for qualitative content analysis (representing the most connected papers for each cluster). Additionally, we selected the 10% papers with the highest weighted degree in the main graph for content analysis (i.e., representing the most connected papers over all the clusters, i.e., in the whole network). We decided to use the weighted degree as a measure for centrality, because our goal was to identify the most representative and strongest connected nodes in each cluster.

3.4. Content analysis

To analyze the content of our literature network and the respective clusters, we chose the *scoping review* approach. It is defined as a systematic process to map existing literature on a

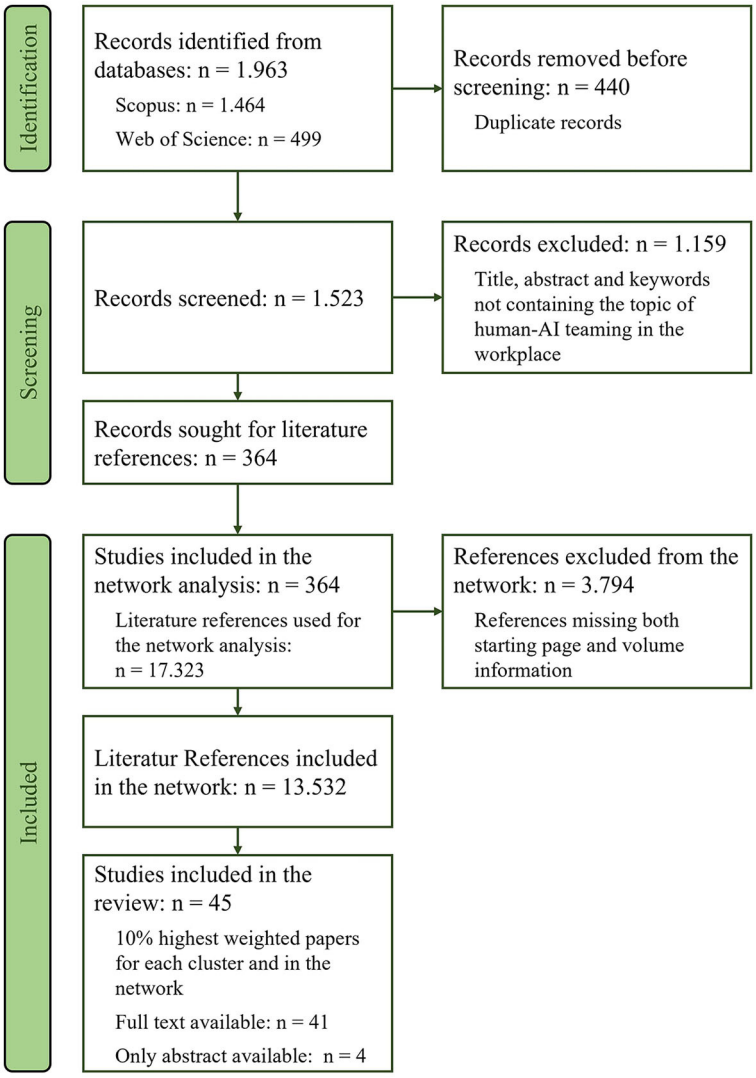

FIGURE 2
PRISMA-diagram of the conducted literature screening process.

research object with the distinctiveness of including all kinds of literature with relevance to the topic, not only empirical work (Arksey and O'Malley, 2005). It is especially of use with emerging topics and evolving research questions (Armstrong et al., 2011) and to identify or describe certain concepts (Munn et al., 2018). Its aims are to show which evidence is present, clarify concepts, how they are defined and what their characteristics are, explore research methods and find knowledge (Munn et al., 2018)—and thus, match our research goals. Whilst this approach lead our systematic literature selection, as described before, it also was our guideline in analyzing the content of the network and the selected publications within.

To understand the network that the respective analysis produced, we looked at the 10% publications with the highest weighted degree in each cluster, analyzing both the metadata such as authors and journals involved, and the content of those papers. For this, we read the full texts of all those publications that were available to us ($n = 41$), as well as the abstracts of the literature without full-text access ($n = 4$). To find the literature's full texts, we looked into the databases and journals that were available to us as university members as well as for open access publication websites, e.g., on Research Gate. For those articles we could not find initially, we contacted the authors. Nevertheless, we could still not get access to four papers, namely Jiang et al. (2022) (cluster 1), Silva et al. (2022) (cluster 3), Tsai et al. (2022) (central within network), Zhang and Amos (2023) (central within network). For those, as they were amongst the most connected publications based on the bibliographic clustering, we considered at least information from the title and abstract.

We first synthesized the main topics of each of the clusters, identifying a common sense or connecting elements within. To then differentiate the clusters, we described them based on standardized categories including the perspective of the articles, research methods used, forms of AI described, role and understanding of AI, terms for and understandings of HAIT and contexts under examination. This, in addition to the

network analysis itself, helped to answer RQ1 on clusters within interdisciplinary HAIT research.

The focus then was on answering RQ2 about the understandings of HAIT represented within the network. For this, we read the full texts central within the clusters and within the whole network, marking all phrases describing, defining, or giving terms for HAIT, presenting the results on a descriptive base. We as well-sorted the network-related papers by the terms they used and the degree of conceptuality behind the constructs to get an idea of terminology across the network.

To answer RQ3 about antecedents and outcomes connected to HAIT, we marked all passages in the literature naming or giving information about antecedents and outcomes. Under antecedents, we understood those variables that have been shown to be preconditions for a successful (or unsuccessful) HAIT. We included those variables that were discussed or investigated by the respective authors as preceding or being needed for teaming (experience), without having a pre-defined model of antecedents and outcomes in mind. For the outcomes, we summarized the variables that have been found to be affected by the implementation of HAIT in terms of the human and technical part, team and task level, performance, and context. We only looked at those variables that were under examination empirically or centrally discussed within the non-empirical publications. Antecedents or outcomes only named in the introductions or theoretical background were not included, as those did not appear vital within the literature. We synthesized the insights for all clusters and gave an overview over all antecedents and outcomes, quantifying their appearance. This was done by listing each publication's individual variables and then subsequently grouping and sorting the variables within our researcher team to achieve a differentiated, yet abstracted picture about all factors under examination within the field of HAIT.

4. Results

4.1. Literature network on human-AI teaming

After removing isolated nodes ($n = 63$) without connections and two articles with missing reference meta-data, the network consisted of 299 nodes (i.e., papers) and 2,607 edges (i.e., paths between the publications). Each paper had on average 17.44 edges connected to it. This is in line with the expected network structure, given that a well-defined and curated part of the literature was analyzed, where most papers share references with other papers. The strength (corrected mean strength = 18.23) was slightly higher than the average degree (17.44), showing a small increase in information gained by using a weighted network instead of an unweighted one. The uncorrected mean strength was 200.55. Transitivity, also known as global clustering coefficient, measures the tendency of nodes to cluster together and can range between the 0 and 1, with larger numbers indicating greater interconnectedness (Ebadi and Schiffauerova, 2015). The observed transitivity was 0.36, which is much higher than random degree of clustering, compared to a transitivity of 0.06 of a random graph with the same number of edges and nodes. The network diameter (longest path between two nodes) was 6, and the density (number of possible vs. observed edges) was 0.06. Overall, this shows that the papers analyzed are part of a connected network that also displays clustering, providing further insights about the network's character.

In total, multilevel community clustering identified five clusters that fit our criteria of a cluster size of ≥20 edges (see Figure 3). The sizes for the five clusters were: $n_1 = 55$, $n_2 = 58$, $n_3 = 55$, $n_4 = 75$, $n_5 = 54$. Thus, all except two edges could be grouped in these clusters. The modularity of the found cluster solution was 0.36. Modularity is a measure introduced by Newman and Girvan (2004) that describes the quality of a clustering solution A modularity of 0 indicates no better clustering solution than random, while the maximum value of 1 indicates a very strong clustering solution. Our observed modularity of 0.36 fell in the lower range of commonly observed modularity measures of 0.3–0.7 (Newman and Girvan, 2004).

4.2. Authors and publication organs within the network

Overall, the network involved about 1,400 authors (including the editors of conference proceedings and anthologies). While most of them were the authors of one to two publications within the network, some stood out with four or more publications: Jonathan Cagan (five papers), Nathan J. McNeese (eight papers) & Beau G. Schelble (four papers), Andre Ponomarev (four papers), Myrthe L. Tielman (four papers) and Dakuo Wang (four papers; see Table 2).

Looking at publication organs, we list all journals, conference proceedings or anthologies of the respective 10% most connected publications within and across the clusters in Figure 4 for economic reasons. To give further insights, we classified those publication organs according to their thematic focus based on color coding.

4.3. Description of clusters within the network

For the content analysis, we decided to include the publications with the 10% highest weighted degree from each cluster to deduce the focus in terms of content and research of these identified clusters and in general. Thus, we read six representative contributions for clusters 1, 2, 3, and 5, eight publications from the larger cluster 4, and for the 10% of articles with the highest weighted degree across the network, another 13 publications were screened, resulting in n = 45 publications within the network being reviewed concerning the topic of human-AI teaming.

Regarding RQ1, we subsequently provide a description of the thematic focus within the five clusters. However, it should be acknowledged that the content of theses clusters exhibits a high degree of interconnectedness, making it more challenging to distinguish between them as originally anticipated. The distinctions among the clusters are based on subtle variations in research orientation or the specific AI systems under investigation. A noteworthy commonality across all clusters is the prevailing technical orientation observed in current HAIT research. This orientation is also reflected in the disciplinary backgrounds of the

researchers involved, with a predominant presence of computer science and engineering expertise across the clusters and in the whole network and partially in the publication organs. The nuanced aspects of this predominantly one-sided perspective, which we were able to discern, are outlined in the subsequent section. Table 2 provides information on the composition of each cluster, including the contributing researchers and the weighted degree of each contribution.

4.3.1. Cluster 1: human-oriented

The 10% most central articles within this first cluster were all journal articles, mostly from ergonomics and psychology-oriented journals: Three of them belonged to Computers in Human Behavior, while the others were from Human Factors, Ergonomics and Information System Frontiers. Two articles shared the two authors McNeese and Schelble. The papers are not regionally focused, with contributions from the US, Germany, Australia, China and Canada. All take a human-oriented approach to HAIT, looking at or discussing a number of subjective outcomes of HAIT such as human preferences, trust and situation awareness. All the papers seem to follow the goal of finding key influencing factors on the human side for acceptance and willingness to team up with an AI. One exception was the paper by O'Neill et al. (2022), which is based more on the traditional technology-centered LOA model in its argumentation, but still reports on many studies looking at human-centered variables.

4.3.2. Cluster 2: task-oriented AI modes

Whilst the 10% most central articles did not have much in common considering geographic origin, authors, journals and

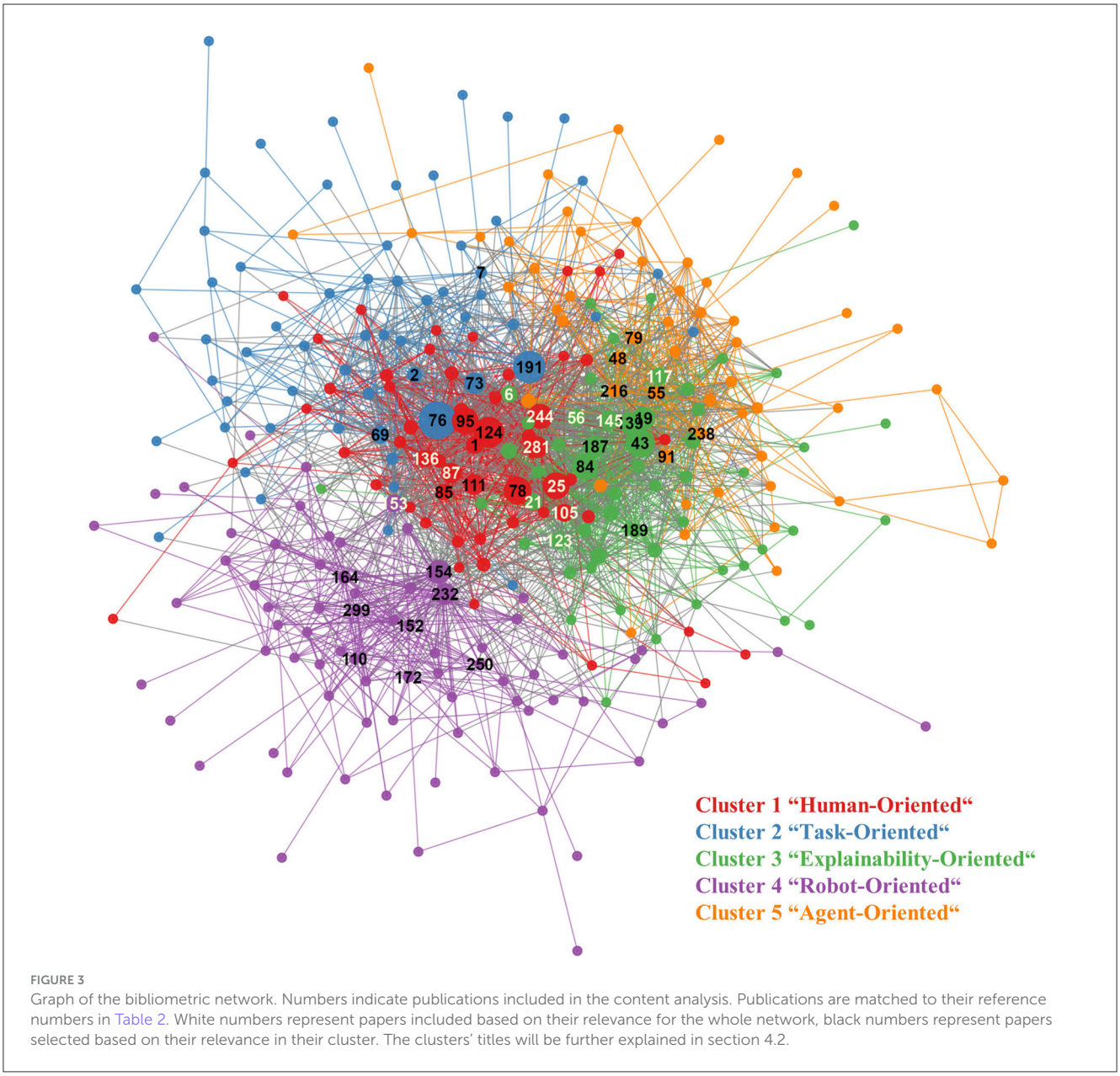

FIGURE 3
Graph of the bibliometric network. Numbers indicate publications included in the content analysis. Publications are matched to their reference numbers in Table 2. White numbers represent papers included based on their relevance for the whole network, black numbers represent papers selected based on their relevance in their cluster. The clusters' titles will be further explained in section 4.2.

TABLE 2 Composition of the identified clusters and strengths (str.) of the included paper.

Cluster 1			Cluster 2			Cluster 3			Cluster 4			Cluster 5			Network's most connected papers		
No.	References	Str.	No.	References	Str.	No.	References	Str.	No.	References	Str.	No.	References	Str.	No.	References	Str. tot.
124	Vössing et al. (2022)	435	76	Yam et al. (2023)	263	43	Fan et al. (2022)	475	232	Castro et al. (2021)	473	216	Kraus et al. (2021)	254	25	Cabour et al. (2022)	731
95	Xiong et al. (2023)	414	69	Jain et al. (2021)	261	84	Naiseh et al. (2023)	428	154	Mukherjee et al. (2022)	418	48	Chong et al. (2023)	194	244	Weisz et al. (2021)	697
111	O'Neill et al. (2022)	393	191	Jain et al. (2022)	261	187	Silva et al. (2022)	423	164	Rodrigues et al. (2023)	310	79	Chong et al. (2023)	174	281	Johnson et al. (2021)	640
78	Endsley (2023)	388	7	Chandel and Sharma (2023)	194	19	Lai et al. (2022)	379	299	Galin and Meshcheryakov (2021)	224	91	Kridalukmana et al. (2022)	172	87	Le et al. (2023)	599
85	Hauptman et al. (2023)	370	73	Jiang et al. (2022)	183	139	Rastogi et al. (2022)	377	172	Othman and Yang (2022)	192	55	Demir et al. (2021)	168	145	Chen et al. (2022)	590
1	Saßmannshausen et al. (2021)	348	2	Li et al. (2022)	164	189	Tabrez et al. (2022)	353	152	Semeraro et al. (2022)	191	238	Wang et al. (2021)	167	105	Verhagen et al. (2022)	534
									110	Dahl et al. (2022)	189				153	Tsai et al. (2022)	524
									250	Aliev and Antonelli (2021)	184				136	Arslan et al. (2022)	523
															56	Cabitza et al. (2021)	517
															6	Zhang and Amos (2023)	517
															117	Fogliato et al. (2022)	511
															21	Pynadath et al. (2022)	509
															123	Cruz et al. (2021)	489

str., node strength based on the cluster's subgraph; str. tot., node strength based on the main graph; no., number of publications within cluster (see Figure 4).

10.3389/frai.2023.1250725

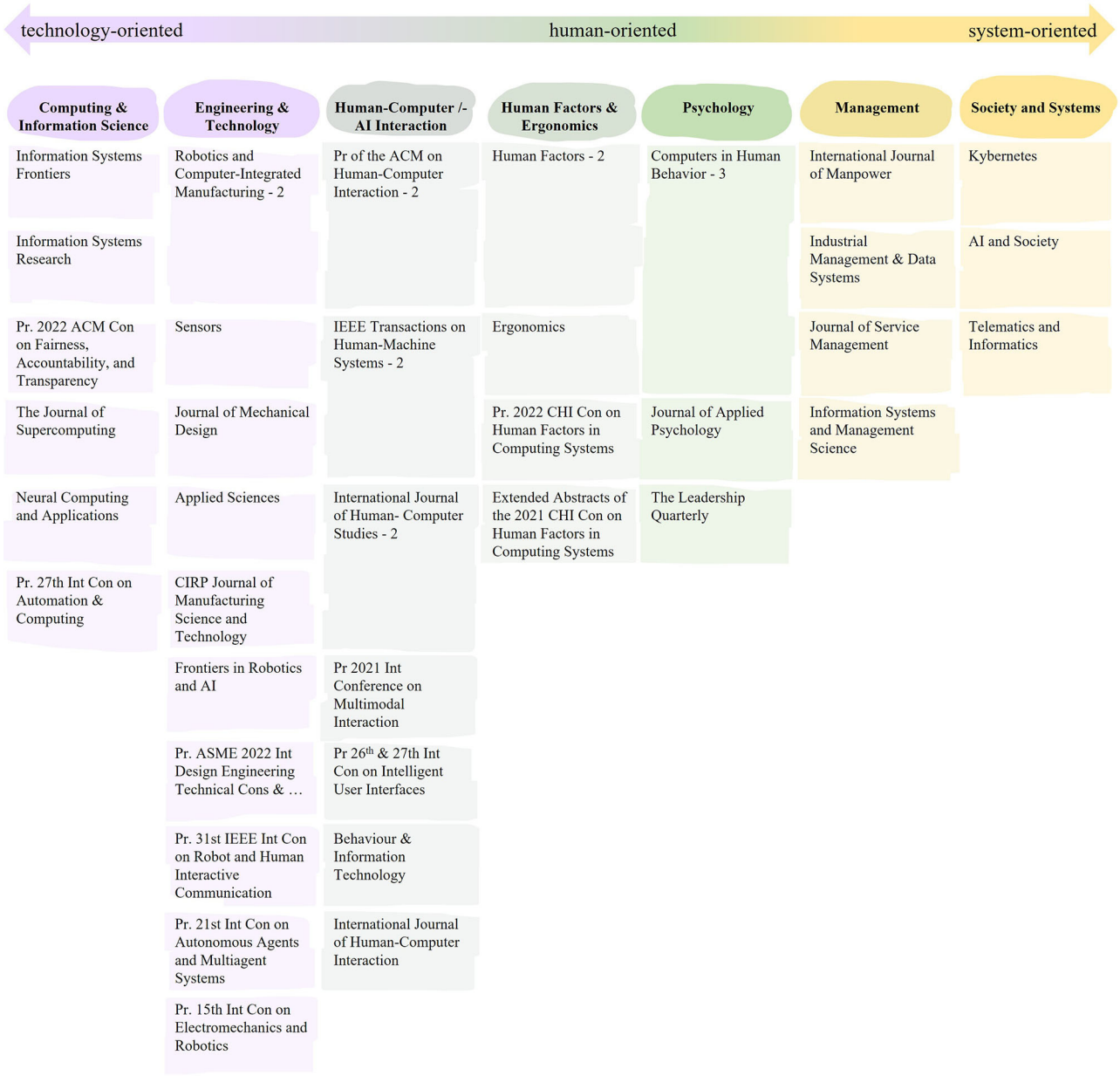

FIGURE 4
Publication organs within the analyzed papers in the network, sorted by their point of orientation. Pr., proceedings; Con, conference; Int., international.

conferences, they share a rooting in information science. All the papers, except for Yam et al. (2023), discuss different types of intelligence automation or roles of AI. They argued from a task perspective, with a focus on the application context and specific ideas for collaboration strategies dependent on the task at hand.

4.3.3. Cluster 3: explainability

The 10% most central articles from Cluster 3 were conference proceedings (four) and journal articles (two) all within the field of human-computer interaction. Three of the articles incorporated practitioner cooperations (with practitioners from Microsoft, Amazon, IBM and/or Twitter). The authors were mainly from the USA, the UK and Canada. Methodologically, the articles were homogeneous in that they all reported laboratory experiments in which a human was tasked with a decision-making scenario during which they were assisted by an AI. The articles took a technical approach to the question of how collaboration, calibrated trust and decision-making can be reached through AI explainability (e.g., local or global explanations, visualizations). Explainability can be defined as an explainer giving a corpus of information to an addressee that enables the latter to understand the system in a certain context (Chazette

et al., 2021). The goal of the articles was to facilitate humans to adequately accept or reject AI recommendations based on the explainability of the system. AI has been characterized as an advisor/helper or assistant and the understanding of AI is focused on the *algorithm/machine-in-the-loop paradigm*, involving algorithmic recommendation systems that inform humans in their judgements. This is seen as a fundamental shift from full automation toward collaborative decision-making that supports rather than replaces workers.

4.3.4. Cluster 4: technology-oriented

Cluster 4 can be described as a *technology-oriented cluster*, which focused primarily on robots as the technology under study. Of the 10% most central articles in this cluster, a majority were journal articles (six), added by two conference contributions. The papers were mainly related to computer science and engineering and similar in their methods, as most of the papers (six) provided literature and theoretical reviews. No similarities could be found regarding the location of publication: While a large part of the articles included in Cluster 4 were published in Europe (Portugal, Scotland, UK, Sweden, and Italy), there were also contributions from Canada, Brazil, and Russia. All included papers dealt with human-robot collaboration as a specific, embodied form of AI, with an overarching focus on the security aspects during this collaboration. The goal of the incorporated studies was to identify factors that are important for a successful collaboration in a modern human-robot collaboration. In this context, communication emerged as an important influencing component, taking place also on a physical level in the case of embodied agents, which necessitates special consideration of security aspects. Furthermore, the articles had a rather technology-oriented approach to safety aspects in common and in most of the articles, concrete suggestions for the development and application of robot perception systems were made. Nevertheless, the papers also discussed the importance of taking human aspects into account in this specific form of collaboration. Additionally, they shared a common understanding of the robot as a collaborative team partner whose cooperation with humans goes beyond simple interaction.

4.3.5. Cluster 5: agent-oriented

The 10% most connected articles within the cluster consisted of conference proceedings (five) and one journal article, all from the fields of human-machine systems and engineering. The authors were mostly from the USA, but also from Germany, Australia, Japan, China and Indonesia and from the field of technology/engineering or psychology. Methodologically, the papers all reported on laboratory or online experiments/simulations. A connecting element between the articles was the exploration of how human trust and confidence in AI is formed based on AI performance/failure. One exception is the paper by Wang et al. (2021), which is a panel invitation on the topic of designing human-AI collaboration. Although it announced a discussion on a broader set of design issues for effective human-AI collaboration, it also addressed the question of AI failure and human trust in AI. In general, the articles postulated that with increasing intelligence, autonomous machines will become teammates rather than tools and should thus be seen as collaboration partners and social actors in human-AI collaborative tasks. The goal of the articles was to investigate how the technical accuracy of AI affects human perceptions of AI and performance outcomes.

The main focus of the clusters, similarities as well as differences are summarized in Table 3. Taken together, the description of the individual clusters reveals slightly different streams of current research on HAIT and related constructs, within the scope of more technology-driven research yet interested in the interaction with humans.

4.4. Understandings of human-AI teaming

To answer RQ2 on understandings of human-AI teaming and to find patterns in terminology and definitions potentially relevant for the research question on a common ground definition, the following section deals with the understandings of human-AI teams that emerged from the individual clusters and the overarching 10% highly weighted papers.

Within *cluster 1*, there were several definitions and defining phrases in the papers. The most prominent and elaborate within the cluster might be that of O'Neill et al. (2022), underlining that "If [the AI systems] are not recognized by humans as team members, there is no HAT" (p. 907) and defining human-autonomy teaming as "interdependence in activity and outcomes involving one or more humans and one or more autonomous agents, wherein each human and autonomous agent is recognized as a unique team member occupying a distinct role on the team, and in which the members strive to achieve a common goal as a collective" (p. 911). This definition is also referred to by McNeese et al. (2021). To this, the latter added the aspects of dynamic adaptation and changing task responsibility. Endsley (2023) differentiated two different views on human-AI work: one being a supportive AI enhancing human performance (which is more of where Saßmannshausen et al., 2021 and Vössing et al., 2022 position themselves), and one being human-autonomy teams with mutual support and adaptivity (thereby referring to the National Academies of Sciences, Engineering, and Medicine, 2021). What unites those papers' definitions of HAIT are the interdependency, the autonomy of the AI, a shared goal, and dynamic adaptation.

In *cluster 2*, there were not many explicit definitions of HAIT, but a number of terms used to describe it, with "teaming" not being of vital relevance. Overall, the understanding of HAIT—or cooperation—is very differentiated in this cluster, with multiple papers acknowledging that "various modes of cooperation between humans and AI emerge" (Li et al., 2022, p. 1), comparable to when humans cooperate. The focus in these papers lies on acknowledging and describing those differences. Jain et al. (2022) pointed out that there can be different configurations in the division of labor, dependent on work design, "with differences in the nature of interdependence being parallel or sequential, along with or without the presence of specialization" (p. 1). Li et al. (2022) differentiated between inter- and independent behaviors based on cooperation theory (Deutsch, 1949), describing how the preference for those can be dependent on the task goal. Having this differentiation in mind, intelligence augmentation could happen in different modes

TABLE 3 Description of the clusters.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Perspective	Human-oriented	Task-oriented	Explainability-oriented	Robot-oriented	Agent-oriented
Methods	Mainly mixed methods, qualitative interviews, field and online experiments, literature review	Mixed methods, vignette study, theory and framework development, literature synthetization, commentary, experiment, experience sampling	Mixed-methods, laboratory experiments with Wizard of OZ or real AI	Mainly theory and framework development, literature review, partly enriched with exemplary studies	Laboratory and online experiments, panel invitation
Forms of AI	Decision (support) system, variety of software or embodied agents	Decision (support) system, robot	Decision (support) system, virtual drone	Robots with machine or reinforcement learning techniques	Embodied agents, software
Role and understanding of AI	Different roles from decision support to mutually supporting team member, augmentor of intelligence, support in decision-making, full, active member with an own role, equal partner, social counterpart	Different roles, augmentor of intelligence, decision agent, independent, active agent, partner & teammate	Assistant & helper, advisor	Autonomous agent, (physical) interaction partner	Autonomous agent, conversational agent, partner, teammate rather than tool
Terms for HAIT	Cooperation, human-AI collaboration, human-autonomy-teaming, human-machine team, human-AI teaming	Human-AI cooperation, collaboration, augmented intelligence, human-computer symbiosis	Human-AI collaboration, collaborative partnership, algorithm-in-the-loop, AI-assisted decision-making, human-AI partnership, human-agent/robot/drone teaming	Human-robot interaction, human-robot collaboration, duality, human-robot team	Mixed-initiative interactions, human-AI collaboration, human-AI teaming, autonomy as teammate
Understanding of HAIT	Independent agents working toward a common goal, adaptive roles within team	Differentiated understanding from independent to interdependent, integrated architecture	Spectrum from full automation to full human agency, AI assistance to support humans	Supportive relationship, working together for task accomplishment, co-working with influence of each's values and broadening individual capabilities	Complementary strengths, prompt interaction in response to communication flow, adaptation toward dynamic task to achieve common goal
Contexts under examination	Hospitality, production management, cyber incident response, sequential risky, decision-making, context-free	Context-aware services, managerial decision-making, financial markets, gig economy platforms, autonomous driving	Clinical decision making, user experience, content moderation, performance prediction, gaming	Manufacturing, production, industry, safety, context-free	Design, military, autonomous driving, context-free

or by different strategies, as well as mutually, with AI augmenting human or humans augmenting AI (Jain et al., 2021). This led to different roles evolving for humans and robots, although the distinct, active role of AI was underlined as a prerequisite for teaming (Li et al., 2022; Chandel and Sharma, 2023). The authors claimed that research is needed on the different cooperation modes.

In *cluster 3*, the central papers argued that the pursuit of complete AI automation is changing toward the goal of no longer aspiring to replace domain workers, but that AI "should be used to support" their decisions and tasks (Fan et al., 2022, p. 4) by leveraging existing explainability approaches. In that, the aspiration to reach *collaborative processes* between humans and AI was understood as a "step back" from full automation, which becomes necessary due to ethical, legal or safety reasons (e.g., Lai et al., 2022). Collaboration, along with explainability, is a central topic in cluster 3, which Naiseh et al. (2023, p. 1) broadly defined as "human decision-makers and [...] AI system working together". The goal of human-AI collaboration was defined as "'complementary performance' (i.e., human + AI > AI and human + AI > human)" (Lai et al., 2022, p. 3), which should be reached by explainability or "algorithm-in-the-loop" designs, i.e., a paradigm in which "AI performs an assistive role by providing prediction or recommendation, while the human decision maker makes the final call" (Lai et al., 2022, p. 3). Thus, the understanding of human-AI teaming was based on the perspective that AI should serve humans as an "assistant" (Fan et al., 2022; Lai et al., 2022; Tabrez et al., 2022) or "helper" (Rastogi et al., 2022); the notion of AI being a "team member"

was only used peripherally in the cluster and HAIT was not explicitly defined as a central concept by the selected papers of cluster 3.

In *cluster 4*, which focused mainly on robots as technological implementations of AI, the term teaming was not used once to describe the way humans and AI (or humans and robots) work together. The terms "human-robot interaction" (HRI) and "human-robot collaboration" (HRC) were used much more frequently, with a similar understanding throughout the cluster: An interaction was described as "any kind of action that involves another human being or robot" (Castro et al., 2021, p. 5), where the actual "connection [of both parties] is limited" (Othman and Yang, 2022, p. 1). Collaboration, instead, was understood as "a human and a robot becom[ing] partners [and] reinforcing [each other]" (Galin and Meshcheryakov, 2021, p. 176) in accomplishing work and working toward a shared goal (Mukherjee et al., 2022). Thus, the understanding of collaboration in cluster 4 is similar to the understanding in Cluster 3, differentiating between distinct roles in collaboration as in Cluster 2. The roles that were distinguished in this cluster are the human as a (a) supervisor, (b) subordinate part or (c) peer of the robot (Othman and Yang, 2022). A unique property of cluster 4 involved collaboration that could occur through explicit physical contact or also in a contactless, information-based manner (Mukherjee et al., 2022). The authors shared the understanding that "collaboration [is] one particular case of interaction" (Castro et al., 2021, p. 5; Othman and Yang, 2022) and that this type of interaction will become even more relevant in the future, aiming to "perceive the [technology] as a full-fledged partner" (Galin and Meshcheryakov, 2021, p. 183). However, more research on human-related variables would be needed to implement this in what has been largely a technology-dominated research area (Semeraro et al., 2022).

In *cluster 5*, the understanding of HAIT is based on the central argument that advancing technology means that AI is no longer just a "tool" but, due to anthropomorphic design and intelligent functions, becomes an "effective and empowering" team member (Chong et al., 2023, p. 2) and thus a "social actor" (Kraus et al., 2021, p. 131). The understanding of AI as a team member was only critically reflected in the invitation to the panel discussion by Wang et al. (2021) who mentioned potential "pseudo-collaboration" and raised the question of whether the view of AI as a team member is actually the most helpful perspective for designing AI systems. The shift from automation to autonomy has been stressed as a prerequisite for effective teaming. Thus, rather than understanding HAIT as a step back from full automation (see cluster 3), incorporating autonomous agents as teammates into collaborative decision-making tasks was seen as the desirable end goal that becomes realistic due to technological progress.

In addition to the clusters and their interpretation of teaming, we looked at the *10% papers with the highest weighted degree in the whole network*, i.e., the papers that had the most central reference lists across all the literature on HAIT. We expected those papers to deliver some "common sense" about the core topic of our research, as they are central within the network and connected with papers from all clusters. Contrary to our expectations, none of those articles focused on trying to classify and differentiate the concept of HAIT from other existing terminologies in order to create a common understanding across disciplines. See Figure 5 for a classification of the articles based on the extent to which the construct was defined in relation to the term used to depict collaboration.

Four of the central papers showed attempts to define HAIT or related constructs: In the context of human-robot teaming, Verhagen et al. (2022) explored the concept of HART (human-agent/robot team), which encompassed the collaboration and coordination between humans and robots in joint activities, either acting independently or in a synchronized manner. A key aspect emphasized by the authors is the need for mutual trust and understanding within human-robot teams. Similarly, the study conducted by Le et al. (2023) also used robots as interaction partners, although the terminology used was "collaboration". They drew a comparison between the streams of research focusing on human-robot collaboration, which is technically oriented, and human-human collaboration, which is design oriented. To develop their approach to human-robot collaboration, they considered not only the relevant literature on collaboration, but also the theory of interdependence (Thibaut and Kelley, 1959). In turn, Johnson et al. (2021) discussed the concept of human-autonomy teaming and emphasized the importance of communication, coordination, and trust at the team level, similar to Verhagen et al. (2022). Their perspective was consistent with the traditional understanding of teaming, recognizing these elements as critical factors for successful teamwork. Another perspective was taken by Cabitza et al. (2021) who used the term "interaction" to a large extent including AI not only for dyadic interaction with humans but also as a supportive tool for human decision teams. They emphasized a contrast to the conventional understanding of human-AI interaction, which views AI either as a tool or as an autonomous agent capable of replacing humans (Cabitza et al., 2021).

The remaining papers referred to HAIT or related constructs in their work but provided minimal to no definition or references for their understanding: Arslan et al. (2022) emphasized that AI technologies are evolving "beyond their role as just tool[s]" (Arslan et al., 2022, p. 77) and are becoming visible players in their own right. They primarily used the term "interaction" and occasionally "collaboration", focusing on the team level without delving into the characteristics and processes of actual teaming. Cabour et al. (2022), similar to Cruz et al. (2021), discussed HAIT only within the context of explainable AI, without providing a detailed definition or explanation. Cruz et al. (2021) specifically used the term "human-robot interaction" rather than teaming, where the robot provides explanations of its actions to a human who is not directly involved in the task. Emphasizing the "dynamic experience" (Chen et al., 2022, p. 549) of both parties adapting to each other, Chen et al. (2022) used mostly the term "human-AI collaboration". They adopted a human-centered perspective on AI and the development of collaboration. In addition, the paper by Tsai et al. (2022) discussed human-robot work, primarily using the notion of collaboration to explore different roles that robots can take, including follower, partner, or leader. The paper by Zhang and Amos (2023) focused on collaboration between humans and algorithms. Fogliato et al. (2022) focused on "AI-assisted decision-making" (p. 1362) and used mainly the term "collaboration" to describe the form of interaction. They only used the term "team" to

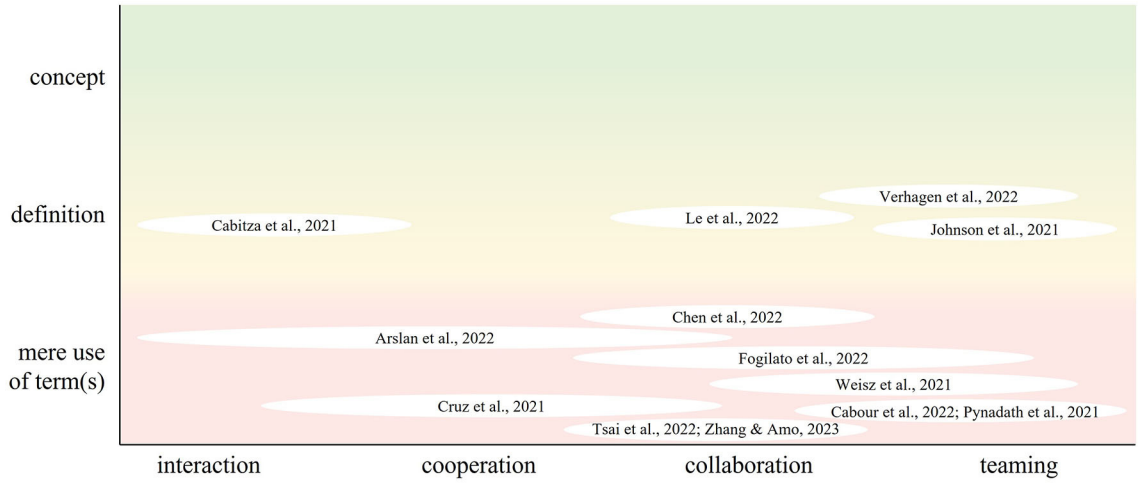

FIGURE 5
Papers with the most impactful connections within the network on HAIT, classified according to their definitory approach and their use of terms for teaming. "Mere use of term(s)" refers to using one of the listed terms without employing or referencing a definition. "Definition" includes the articles in which the understanding of the used teaming term is specified. "Concept" refers to a deep understanding toward the used term, e.g., by differentiating it from other terms or deriving/proposing a definition.

describe the joint performance output without further elaboration on its characteristics or processes. Weisz et al. (2021) took the notion of teaming a step further, discussing future potential of generative AI as a collaborative partner or teammate for human software engineers. They used terms such as "partnership," "team," and "collaboration" to describe the collaborative nature of AI working alongside human engineers. Finally, Pynadath et al. (2022) discussed human-robot teams and emphasized the "synergistic relationship" (p. 749) between robots and humans. However, they also did not provide additional explanations or background information on their understanding of teamwork.

What we see overall is that there are different streams of current research on HAIT, examining different aspects or contexts of HAIT. Whilst there is one cluster centered around human perception of HAIT, with a tendency to use the term teaming, the other clusters focus more on the AI technology or on the task, describing teaming in a sense of cooperation or collaboration, partially envisioning the AI as a supportive element. Also, within the network's most connected papers, we find this diversity in understandings and terminology and, yet again, a lack in conceptual approaches and definitions.

4.5. Antecedents and outcomes

To structure the antecedents and outcomes under examination within the clusters on RQ3, we developed a structural framework helping to group them according to the part of the (work) system they refer to. We used the structuring of Saßmannshausen et al. (2021) as an orientation, who differentiate AI characteristics, human characteristics and (decision) situation characteristics as categories for antecedents. As our reference was HAIT and not only the technology part (as with Saßmannshausen et al., 2021), we needed to broaden this scheme and chose the categories of human, AI, team, task (and performance for outcomes) and context to describe the whole sociotechnical system. We as well-added a perception category for each category to clearly distinguish between objectively given inputs (see also O'Neill et al., 2022) and their subjective experience, both being potential (and independent) influence factors or outcomes of HAIT. Note that all antecedents and outcomes were classified as such by the authors of the respective publications (e.g., by stating that "X is needed to form a successful team") and can relate to either building a team, being successful as a team, creating a feeling of team cohesion etc. The concrete point of reference differs depending on the publication's focus but is always related to teaming of human and AI.

Cluster 1 contained a high number of antecedents of HAIT or variables necessary to it such as trust. Amongst these were the (dynamic) autonomy of the AI, trust, but also aspects relating to explainability of the AI and situation awareness. Two of the papers took a more systematic view on antecedents, structuring them into categories. The review by O'Neill et al. (2022) contained in this cluster, sorts the antecedents they found into characteristics of the autonomous agents, team composition, task characteristics, individual human variables and training. Communication was found to serve as a mediator. Saßmannshausen et al. (2021) structure their researched antecedents (of trust in the AI team partner) into AI characteristics, human characteristics and decision situation characteristics. For outcomes, cluster 1 included—next to a number of performance- and behavioral outcomes—many different subjective outcomes, e.g., perceptions of the AI characteristics, perceived decision authority, mental workload or willingness to collaborate. O'Neill et al. (2022) did not provide empirical data on outcomes of HAIT itself, but presented an overview of the literature on various outcomes, including

performance on the individual and team level (70 studies), workload (39 studies), trust (24 studies), situation awareness (23 studies), team coordination (15 studies) and shared mental models (six studies).

Cluster 2 incorporated relatively few antecedents and outcomes of teaming, as most papers focused on the structure or mode of teaming itself. These cooperation modes could be considered as the central antecedent of the cluster. AI design, explainability as well as the specificity of the occupation, task (and goal) or the organizational context were also named. They were supposed to affect subjective variables such as trust, role clarity, attitude toward cooperation and preference for a feedback style, but also broad organizational aspects such as competitive advantages.

In cluster 3, AI explainability emerged as the main antecedent considered by all central articles. The articles differed in the way that explainability was technologically implemented (e.g., local vs. global explanations), but all considered it as an antecedent for explaining outcomes related to calibrated decision-making (objective, i.e., accuracy of decisions as well as subjective, i.e., confidence/trust in decision).

The majority of the contributions in cluster 4 consisted of theoretical reviews and frameworks, in which antecedents of a successful human-robot collaboration were derived and discussed. Identified antecedents, primarily related to the physical component of a robotic system, were robot speed, end-effector force/torque, and operational safety aspects. Indicated antecedents, which were discussed and can also be applied to non-embodied AI systems, were the ability of the system to learn and thus to generalize knowledge and apply it to new situations, as well as effective communication between the cooperation partners, a shared mental model to be able to work toward the same goal, and (bidirectional) trust. In addition, the usability of the system, its adaptability, and ease of programming, the consideration of the psychophysiological state of the human (e.g., fatigue, stress) and the existing roles in the workplace were identified as prerequisites for a create harmonious collaboration between humans and technologies. When considering the antecedents addressed, expected outcomes included increased productivity and efficiency in the workplace, reduced costs, and better data management.

The articles in cluster 5 considered or experimentally manipulated AI performance (accuracy, failure, changes in performance) and the general behavior of the system (proactive dialogue). The articles argued that this is a central antecedent for explaining how trust is developed, lost or calibrated in human-AI teams.

Overall, the antecedents and outcomes on HAIT have received a large amount of research interest, thus a number of variables have already been studied in this context (see Tables 4, 5 for an overview).

4.6. Definition of human-AI teaming

Our final RQ4 was to identify, if feasible, a cohesive definition that would bridge the diverse aspects addressed in current HAIT research. However, as evident from the results of the other research questions, a lack of defining approaches and concepts is apparent throughout the network. We only found one elaborate definition with O'Neill et al. (2022), which was also cited, but not by the breadth of publications. Notably, the included publications, including O'Neill et al. (2022), predominantly adopt a perspective that focuses on one of the two subsystems within a team (i.e., the human or the AI), and tend to be primarily technology-oriented. That means that it is mainly examined which conditions a technical system needs for teaming or, which characteristics the human being should bring along and how these can be promoted for collaboration. This one-sided inclination is also reflected in the addressed antecedents and outcomes (see Tables 4, 5).

However, in order to foster a seamless teaming experience and promote effective collaboration, it is crucial to consider the team-level perspective as a primary focus. Questions regarding the requisite qualities for optimal human-AI teams and the means to measure or collect these qualities remain largely unaddressed in the included publications, resulting in a blind spot in the network and the current state of HAIT research, despite the fundamental reliance on the concept of teaming. While the review of O'Neill et al. (2022) on human-autonomy teaming dedicates efforts toward defining the concept and offering insights into their understanding, an extension of this concept, particularly with regard to the team-level perspective, is needed. The subsequent sections of the discussion will expound on the reasons for this need in greater detail and propose an integrative definition that endeavors to unite all relevant perspectives.

5. Discussion

In this work, we aimed to examine the current scope and breadth of literature of HAIT as well as research streams to comprehend the study field, the existing understandings of the term and important antecedents and outcomes. For this purpose, we conducted a bibliometric network analysis revealing five main clusters, followed by a scoping review examining the content and quality of the research field. Before delving into the terminology and understanding of HAIT and what we can conclude from the antecedents and outcomes under examination, we point out the boundaries and connected risks of our work. This serves as the background for our interpretation and the following idea of conceptualizing and defining the construct of HAIT, which is complemented by demands for future research from a perspective on humane work-design and socio-technics.

5.1. Limitations

Choosing our concrete approach of a bibliometric network analysis and follow-up scoping review helped us answer our research questions, despite posing some boundaries on the opportunity of insight. First of all, the chosen methods determined the kind of insights possible. Network analyses rely on citation data to establish connections between publications (Bredahl, 2022). Thereby, the quality and completeness of the citation data may vary, leading to missing or insufficient citations of certain publications, thus causing bias and underrepresentation of certain papers or research directions (Kleminski et al., 2022). We are not aware of

TABLE 4 Antecedents of human-AI teaming.

Category	Antecedent	No of sources	Sources	Cluster(s)
Human	Individual human variables	3	O'Neill et al., 2022; Othman and Yang, 2022; Xiong et al., 2023	1, 4
	Digital affinity, including Aversion to AI	2	Saßmannshausen et al., 2021; Jain et al., 2022	1, 2
	Psychophysiological state	2	Galin and Meshcheryakov, 2021; Mukherjee et al., 2022	4
	Control	1	Vössing et al., 2022	1
	Mental model of situation	1	Endsley, 2023	1
Perception of human	/	0		
AI	Explainability of AI, including Local and global explanations Visualizations/guidance Explanation and information about decision uncertainty Transparency	10	Fan et al., 2022; Kridalukmana et al., 2022; Lai et al., 2022; O'Neill et al., 2022; Rastogi et al., 2022; Tabrez et al., 2022; Vössing et al., 2022; Chandel and Sharma, 2023; Endsley, 2023; Naiseh et al., 2023	1, 2, 3, 5
	Design Minding human cognitive skills and limitations Organization-specific adaptation	3	Jain et al., 2021; Jiang et al., 2022; Vössing et al., 2022	1, 2
	Difficulty of programming	1	Galin and Meshcheryakov, 2021	4
	LOA/AI autonomy Partial vs. full Restrictions in autonomy Proactivity of AI	4	Kraus et al., 2021; Mukherjee et al., 2022; O'Neill et al., 2022; Hauptman et al., 2023	1, 4, 5
	Dynamics AI adaptivity AI adaptability	3	Galin and Meshcheryakov, 2021; Mukherjee et al., 2022; Hauptman et al., 2023	1, 4
	AI performance, including Good vs. bad performance Failures Changes in performance Reliability	3	Demir et al., 2021; O'Neill et al., 2022; Chong et al., 2023	1, 5
	Guaranteed safety of the AI	1	Galin and Meshcheryakov, 2021	4
	Openness to human scrutiny	1	Chandel and Sharma, 2023	2
	Conformability of the AI	1	Galin and Meshcheryakov, 2021	4
Perception of AI	Predictability of AI actions	3	Aliev and Antonelli, 2021; Mukherjee et al., 2022; Hauptman et al., 2023	1, 4
	Perceived AI comprehensibility	1	Saßmannshausen et al., 2021	1
	(Bidirectional) trust, including trusting behavior	4	Saßmannshausen et al., 2021; Mukherjee et al., 2022; Semeraro et al., 2022; Vössing et al., 2022	1, 4
	Perceived AI ability	1	Saßmannshausen et al., 2021	1
Team	Team interaction, including Communication	5	Castro et al., 2021; Demir et al., 2021; Mukherjee et al., 2022; O'Neill et al., 2022; Othman and Yang, 2022	1, 4, 5
	Interdependence between human and AI	1	Li et al., 2022	2
	Human-robot roles	1	Othman and Yang, 2022	4
	Collaboration mode Sequential or parallel task, with or without specialization, AI or human first	1	Jain et al., 2022	2
	Team composition (members)	1	O'Neill et al., 2022	1

(Continued)

TABLE 4 (Continued)

Category	Antecedent	No of sources	Sources	Cluster(s)
	Team experience level	1	Hauptman et al., 2023	1
	Shared mental models		Castro et al., 2021; Mukherjee et al., 2022	4
	Situation awareness (SA) Shared SA Human SA of AI state AI SA on state of human	1	Endsley, 2023	1
Perception of team	/	0		
Task	Task characteristics	2	Mukherjee et al., 2022; O'Neill et al., 2022	1, 4
	Work phase	1	Hauptman et al., 2023	1
	Goal orientation (task)	1	Li et al., 2022	2
	Time for decision making	1	Rastogi et al., 2022	3
Perception of task	Ease of critical information transferring	1	Othman and Yang, 2022	4
Context	Effects of the (joint) decision Probability of significant and irreversible changes	1	Hauptman et al., 2023	1
	Training/learning, including Time needed or used for acceptance and understanding of AI	3	Castro et al., 2021; O'Neill et al., 2022; Hauptman et al., 2023	1, 4
	Type of workspace	1	Mukherjee et al., 2022	4

a bias toward certain journals, geographic regions or disciplines within our network, but do not know if this also holds for the cited literature. This might lead to certain areas of HAIT research, such as literature on the teaming level, not being considered by the broad body of literature or by the most connected papers (maybe also due to the mentioned inconsistent terminology), which would reflect also in the papers' content revealing blind spots. Furthermore, bibliometric network analyses focus mainly on the structural properties of the network and hence often disregard contextual information (Bornmann and Daniel, 2008), which is why we decided to conduct a scoping review additionally. Scoping reviews are characterized by a broad coverage of the research area (Arksey and O'Malley, 2005), which is both a strength and a weakness of the method: On the one hand, a comprehensive picture of the object of investigation emerges, but on the other hand, a limitation in the depth of detail as well as in the transparency of quality becomes apparent. Only being able to look into the 10% most connected papers within each cluster also limited our opportunity to go into more detail and map the whole field of research, again with the risk of leaving blind spots that are actually covered by literature. Hence, we considered also the most connected papers within the whole network to get a broader picture.

The basis of our network analysis and review was a literature search in WoS and Scopus. Although these are the most comprehensive databases available (Kumpulainen and Seppänen, 2022), there is a possibility that some relevant work are not listed there or were not identified by our search and screening strategy. More than in the databases, this problem might lie in restricting our search to publications published from 2021 onwards. It might be that important conceptual and definitory approaches can be found in the prior years, although we found no indications for that within the qualitative analyses of terminology or referenced definitions. Confining our search strongly to the last 2 years of research enabled us to address a relatively wide spectrum of the latest literature in a field that is very hyped and has a large output of articles and conference contributions. While there is a risk associated with excluding "older" research, we sought to partially balance it out by analyzing the papers' content, including their references to older definitions and concepts. Nonetheless, it remains a concern that our conclusions may primarily apply to the very latest research stream, potentially overlooking an influential stream of, for instance, team-level research on HAIT, that held prominence just a year earlier. Therefore, it is important to view our results as representing the latest research streams in HAIT.

Finally, bibliometric studies analyze only the literature of a given topic and time period (Lima and de Assis Carlos Filho, 2019), which can limit our results because of research not being found under the selected search terms, and the clustering algorithms used are based on partially random processes (Yang et al., 2016), which limits transparency on how results are achieved. We tried to balance this out by properly documenting our whole analysis procedure and all decisions taken within the analysis.

Another limitation was discovered in our results during the analyses. Our primary idea was to find different clusters in the body of literature which illuminate the construct HAIT from different disciplinary perspectives. From this, we wanted to extract the, potentially discipline-specific, understandings of HAIT and compare them among the clusters. Although we identified five clusters approaching HAIT with different research foci, they did not differ structurally in their disciplinary orientation. The differences in terminology and understanding within the clusters sometimes were just as high as between. Almost all of the identified publications, as well as most of the clusters, took a

TABLE 5 Outcomes of human-AI teaming.

Category	Outcome	No of sources	Sources	Cluster(s)
Human	Human agency	2	Fan et al., 2022; Tabrez et al., 2022	3
	Preference for feedback	1	Jain et al., 2022	2
Perception of human	Perceived decision authority	1	Xiong et al., 2023	1
	Subjective workload	2	Lai et al., 2022; Xiong et al., 2023	1, 3
	Fatigue	2	Galin and Meshcheryakov, 2021; Semeraro et al., 2022	4
	Stress	2	Galin and Meshcheryakov, 2021; Tabrez et al., 2022	3, 4
	Fear	1	Galin and Meshcheryakov, 2021	4
	Role clarity	1	Jain et al., 2022	2
AI	/	0		
Perception of AI	Trust/confidence in AI	12	Demir et al., 2021; Kraus et al., 2021; Fan et al., 2022; Jain et al., 2022; Kridalukmana et al., 2022; Rastogi et al., 2022; Tabrez et al., 2022; Vössing et al., 2022; Chong et al., 2023; Endsley, 2023; Naiseh et al., 2023; Xiong et al., 2023	1, 2, 3, 5
	Comfort with AI teammate	1	Hauptman et al., 2023	1
	Acceptance of AI/willingness to collaborate As replacement of a human teampartner	3	Li et al., 2022; Chong et al., 2023; Xiong et al., 2023	1, 2, 5
	AI legitimacy as a team member	1	Hauptman et al., 2023	1
	Social presence	1	Fan et al., 2022	3
	Evaluation of AI autonomy	1	Hauptman et al., 2023	1
	Perceived AI capability/understanding of AI	3	Fan et al., 2022; Lai et al., 2022; Xiong et al., 2023	1, 3
	User experience, including Engagement, subjective perception	3	Fan et al., 2022; Lai et al., 2022; Xiong et al., 2023	1, 3
	Satisfaction with AI	2	Kraus et al., 2021; Fan et al., 2022	3, 5
Team	Human-machine augmentation	1	Chandel and Sharma, 2023	2
	Situation awareness	1	Tabrez et al., 2022	3
	Decision style matching	1	Xiong et al., 2023	1
Perception of team	Interaction experience	1	Galin and Meshcheryakov, 2021	4
	Attitude toward collaboration	1	Li et al., 2022	2
	Preference for a collaboration mode	2	Li et al., 2022; Xiong et al., 2023	1, 2
	Responsibility attribution	1	Xiong et al., 2023	1
Task	Data management	1	Othman and Yang, 2022	4
	Human reliance on AI/adjusted decision making	2	Rastogi et al., 2022; Vössing et al., 2022	1, 3
Perception of task	Perception of task interdependence	1	Xiong et al., 2023	1
Performance	Performance Human performance (e.g., time/number of pauses) AI performance (e.g., efficacy, precision) HAIT performance (e.g., quality of decision)	10	Saßmannshausen et al., 2021; Fan et al., 2022; Lai et al., 2022; Rastogi et al., 2022; Tabrez et al., 2022; Vössing et al., 2022; Chong et al., 2023; Endsley, 2023; Naiseh et al., 2023; Xiong et al., 2023	1, 3, 5
	Cost reduction	1	Othman and Yang, 2022	4
Perceived performance	Perception of efficiency increase through AI	1	Othman and Yang, 2022	4

(Continued)

TABLE 5 (Continued)

Category	Outcome	No of sources	Sources	Cluster(s)
	Perception of AI performance	1	Xiong et al., 2023	1
	Human confidence in decisions	2	Lai et al., 2022; Tabrez et al., 2022	3
	Confidence in own performance (human)	1	Chong et al., 2023	5
	Perception of task performance	1	Xiong et al., 2023	1
Context	Perceived risk (of a decision)	1	Xiong et al., 2023	1
	Trust in the team by stakeholders	1	Hauptman et al., 2023	1

more technology-centered perspective, which means that some disciplines are not broadly covered in our work. For example, psychological, legal, societal, and ethical perspectives are poorly represented in our literature network. An explanation for this may be that there has been little research on HAIT from these disciplines, or that publications within the network that were not included in the review on a content base or literature form former years not included in our network highlighted these perspectives. Finally, it should be noted that even though very different aspects are researched and focused on within the clusters, the understanding of the construct of HAIT within which the research takes place is either not addressed in detail or only in very specific aspects, limiting our ability to answer our RQ2 adequately.

5.2. Looking at the results: what we know about HAIT so far

Summarizing the findings within our literature network on HAIT under examination or discussion, we can identify some general trends, but also some research gaps and contradictions.

5.2.1. Current research streams and understandings

To answer *RQ1* about human-AI teaming research clusters, we identified five distinct clusters with varying emphases. Despite their shared focus on technological design while considering human aspects, which also reflects in the network metrics, subtle differences in research foci and the specific AI systems under investigation were discernible: Cluster 1 focuses mainly on human variables that are important for teaming. Cluster 2 examines task-dependent variables. Cluster 3 especially investigates the explainability of AI systems, cluster 4 concentrates on robotic systems as special AI applications, and cluster 5 deals mainly with the effects of AI performance on humans' perception. Except for cluster 1, the publications exhibit a focus on technology and are grounded in engineering principles. This is reflected in the publication organs, which are mainly technically oriented, with many at the intersection of human and AI, but primarily adopting a technological perspective. While other perspectives exist, they are not as prevalent. While reasonable due to technological system development's origin in this field (Picon, 2004), research should allocate equal or even more attention to the human and team component in in socio-technical systems. Human perceptions can impact performance (Yang and Choi, 2014), contrasting with technological systems that perform independently of perceptions and emotions (Šukjurovs et al., 2019). However, current research streams continue to emphasize the technological aspects.

Regarding *RQ2,* both terminologies and their comprehension within the clusters were examined to investigate the understanding of HAIT. A broad range of terms is used, often inconsistently within publications. While "teaming" is occasionally used, broader terms like "interaction" and "cooperation" prevail, with "collaboration" being the most common. Interestingly, many terms used do not focus on the relational or interactional part of teaming but instead highlight technology as support, a partner or a teammate, reflecting the technology-centeredness once again. In parallel, it becomes apparent that the phenomena of work between humans and AI systems are rarely defined or classified by the authors. Instead, the terms "cooperation," "collaboration," "interaction," and "teaming" are used in a taken-for-granted and synonymous manner. Paradoxically, a differentiated understanding emerges in some of the papers: "interaction" denotes shared workspace and task execution with sequential order or just any contact between human and AI, "cooperation" involves access to shared resources to gather task-related information, but retains separate work interests, and "collaboration" entails humans and technologies working together on complex, common tasks. However, this differentiation that is very established in human-robot interaction research (see, e.g., Othman and Yang, 2022), is not consistently reflected within the majority of papers within our network. Except for O'Neill et al.'s (2022) paper, the term "teaming" is underdefined or unclassified in other works. Possible reasons include the dominance of a technology-centric perspective (Semeraro et al., 2022) in current research efforts, as collaboration aspects are likely to attract more interest in other research domains, such as psychology or occupational science (Bütepage and Kragic, 2017). Regarding the exemplary publication organs, those are underrepresented in our network. Another possible reason could be the novelty of the research field of teaming with autonomous agents (McNeese et al., 2021). Compared to the other definable constructs, the concept of teaming

has only been increasingly used in recent years, which means that research in this field is still in its infancy and, thus, it has not yet fully crystallized what the defining aspects of teaming are. However, it raises questions about conducting high-quality research in the absence of a well-defined construct, as terms like "teammate" or "partner" alone lack the scientific clarity required for construct delineation.

One interesting idea shown in some of the publications offers a way to unite the different terms used within the field: the concept of existing collaboration modes or different views on human-AI work. Authors such as McNeese et al. (2021), Li et al. (2022), Chandel and Sharma (2023), and Endsley (2023) address that there might be different ways (or degrees) of AI and humans collaborating: Some aim to support the human, which reflects more of a cooperative perspective with distinct, not necessarily mutually interdependent tasks. Others are conceptualized as human-AI teams from the very beginning, with mutual intelligence augmentation, dynamic adaptation to one another and collaborative task execution. One can discuss if these should be seen as different categories of interaction, or if they are considered different points on a continuum of working together.

5.2.2. Antecedents and outcomes

To answer *RQ3* on antecedents and outcomes of HAIT we note that for antecedents, nearly all components of a human-AI team were under examination or discussion at least in a few publications, except for team and human perception. Research on AI characteristics dominated the field, with many constructs under research from the, apparently most important, topic of explainability (10 publications) to dynamics and levels of automation of AI. For team variables, most papers looked at team interaction as well as the conglomerate of (shared) situation awareness and mental models. What we can see overall is a focus on characteristics of the work system, but also quite a few perceptional and subjective antecedents under investigation. This shows the importance of considering not only objectively given or changeable characteristics, e.g., in AI design, but also how humans interact with those characteristics, how they perceive them on a cognitive and affective level.

For the outcomes, we find that trust (11 publications) and performance (10 publications) are by far the most researched and discussed outcomes of human-AI teaming. This is interesting, as they represent both the objective, economically important side of implementing teams of AI and humans, but also the subjective basis for efficient long-term collaboration. In the studies, we find a strong focus on subjective outcomes, considering the perception of oneself within the work situation (e.g., stress or fear), the perception of the AI (e.g., comfort with it, perceived capabilities) which is a focus of the literature with 26 mentions, and the perception of the team (e.g., preference for a collaboration mode) as well as its performance.

Nevertheless, considering human perception in researching and designing HAIT is only the first step toward reaching human-centeredness. This approach portrays the human as the central role within complex sociotechnical systems (Huchler, 2015). As a research philosophy, it goes beyond measuring trust or including some worker interviews in one's research and understands the human (and, e.g., their trust in an AI system) as the starting point of any system design. This perspective perfectly goes along with other conceptual approaches such as a socio-technical thinking (see, e.g., Emery, 1993) or the idea of Industry 5.0 (Breque et al., 2021). The breadth of different antecedents and outcomes found in the field of literature on HAIT is impressive, showing knowledge on specific aspects on HAIT and an interest in interdisciplinarity and finding out about different aspects preceding or resulting from HAIT. Still, it lacks a conceptual underpinning that is holistically considering the human as the central figure within a work system.

5.2.3. Exploring existing definitions of HAIT

What we can see considering current understandings of human-AI teams is that many of the publications involved some definitory elements, be it the support aspect, shared mental models, or mutual communication, but all were very focused on those (or other) specific aspects. Nearly no publication clearly defined HAIT in their theoretical background as a basis of their work—most publication use it in a way as if it was self-explanatory. Terms for teaming are used inconsistently and differentiations between them are only addressed in some publications on different cooperation modes. However, the range of terminology, as well as the multitude of disciplines and perspectives contributing to the study of HAIT, permit extensive exploration and the generation of numerous fresh insights. This diversity is appropriate for a field of research that is just evolving. Nonetheless, in order to enhance the clarity and cohesiveness of the literature in this field, there is a pressing need for a unified conceptual framework that allows for transparency (Flake and Fried, 2020) and illuminates how the amalgamation of various attributes can effectively shape humans and AI into a team. We were not able to find such a widely accepted, clear and comprehensive definition of HAIT that would fully answer *RQ4*. This is a problem that links back to the research topic of Human-Centered AI at Work and its aim to find common ground in theories and methods. To better answer *RQ4*, we therefore developed an own definition on HAIT, which is derived in section 5.3.3.

5.3. What we need for HAIT: integrated, well-defined teaming approaches

Overall, a great interest in HAIT research can be seen. Studies are being published successively on this topic, being connected through a network of references, and many variables are examined. Some of them are investigated extensively, such as explainability or trust, while there is a variety of variables that is more exploratory examined in single studies. What is lacking, however, is a defined construct that would systematize the understanding toward HAIT and lead to unified and more

integrated research. There is little effort in creating a unified definition for the teaming aspect of humans and AI working together; rather, the focus is still primarily on how to prepare the technological counterpart for collaborating. The way toward a common ground is still to be gone, but our review helps identify what is needed next.

The different terms used, lack of definitions and concepts, and various understandings of what constitutes "teaming" and what role(s) the AI might take make it difficult to unify research, build common ground, and advance the field. Hence, we see the need for…

1. addressing HAIT from a socio-technical perspective, thus strengthening the teaming idea and human-centeredness.
2. understanding the AI as a team partner able to take roles adaptively instead of holder of one specific role.
3. a clear definition and a distinct terminology, that is grounded in the work so far and that has the potential to be referred to and used in future research.

5.3.1. The teaming idea within human-AI teams from a socio-technical perspective

What we have seen throughout the review is the vast interest in human-related variables, that show the importance of a human-centered understanding and a consideration of the whole socio-technical system when examining and designing HAIT. Still, this interest does not yet result in taking a human- or even team-oriented perspective. One of the few definitory approaches of O'Neill et al. (2022), focusses on what the AI needs to be and contribute to enable teaming, and not on what this teaming actually is. Thus, research needs to take a holistic approach involving multiple disciplines to investigate and design functioning, accepted and adaptable collaboration between humans and AI. This idea is not new in itself, but follows the concept of socio-technical system design (see, e.g., Emery, 1993), where work systems are seen as consisting of a social and a technical subsystem, connected by organization. Central to that is the approach of *joint optimization*, meaning to design both systems together and constantly adapt them to one another so that both systems yield positive outcomes (Appelbaum, 1997). The epitome of this thinking is the idea of human-AI teaming. It incorporates the idea of humans (social systems) and AI (technical system) working together, creating synergies and jointly forming something that goes beyond their individual capabilities, and thus a new social system. Hence, we want to underline the importance of bringing the teaming idea, and established theories and empirical research from human-human teaming, into the field of research on human-AI or human-autonomy teaming. In most of the literature, terms underlining the collaborative element such as *partner*, *symbiosis* or *teammate* are used as buzzwords without further explanation or without really understanding humans and AI as a sociotechnical system acting as a team. For a clearly defined field of research, future work should therefore think carefully about which construct (e.g., interaction, teaming) is examined and disclose this understanding to the readers. Furthermore, different constructs should not be used synonymously, as this can lead to a deterioration in the quality of research and confusion.

For us, the term and idea of teaming is still central, being reflected in the vast use of associated terms and the omnipresent idea of a new quality of interaction with a development away from the tool perspective, adaptive behavior, and shared mental models. This evokes the need to unite knowledge on (human) teaming with knowledge on AI and human interaction to go a step further and establish a concept of HAIT that is viable for sustaining research and implementing it into practice.

5.3.2. The "role" of AI within the team

Throughout the papers within our network, we have identified various labels and roles for the AI systems described. While most papers primarily focus on one specific role of AI in their investigations, some, such as Endsley (2023), describe different "concepts of operation" (p. 4) like an AI as an aid to a human supervisor, AI as a collaborator, or AI as an overseer and limiter of human performance. She also mentions roles like "coach, trainer or facilitator" (p. 4). These roles can be described by factors like the nature of the task (e.g., exploration and exploitation, see Li et al., 2022), the level of dependence between AI and human, and specialization (Jain et al., 2022). Jain et al. (2022) distinguish between different "work designs", systematically describing the division of labor between humans and AI in different categories. Beyond the literature screened for our review, there are other papers addressing the systematics of human-AI interaction, such as Gupta and Woolley (2021). One notable example with comprehensive categorization is Dellermann et al. (2021), who differentiate between aspects defining AI-human and human-AI interactions.

From our perspective, what is needed is to use these existing delineations and taxonomies to develop a new concept of AI as a dynamic team member, capable of adaptively changing roles as required. In our understanding, HAIT goes beyond mere cooperation or collaboration alone, but it can encompass elements of both. HAIT entails humans and AI working together on the same tasks and goals, adapting and exchanging roles as needed. Sometimes, this involves separate cooperation, but it can switch the "mode of collaboration" to mutual support or to the AI providing guidance to the human executor. This understanding of HAIT transcending the categories of cooperation and coordination and including a wide range of potential roles for both humans and AI is depictured in Figure 6.

This concept aligns with the idea of augmented intelligence, as described by Jain et al. (2021), where "computers and humans working together, by design, to enhance one another, such that the intelligence of the resulting system improves" (p. 675). Building on the present research and knowledge about specific roles and cooperation modes, the next step in research is a more realistic, dynamic utilization of AI

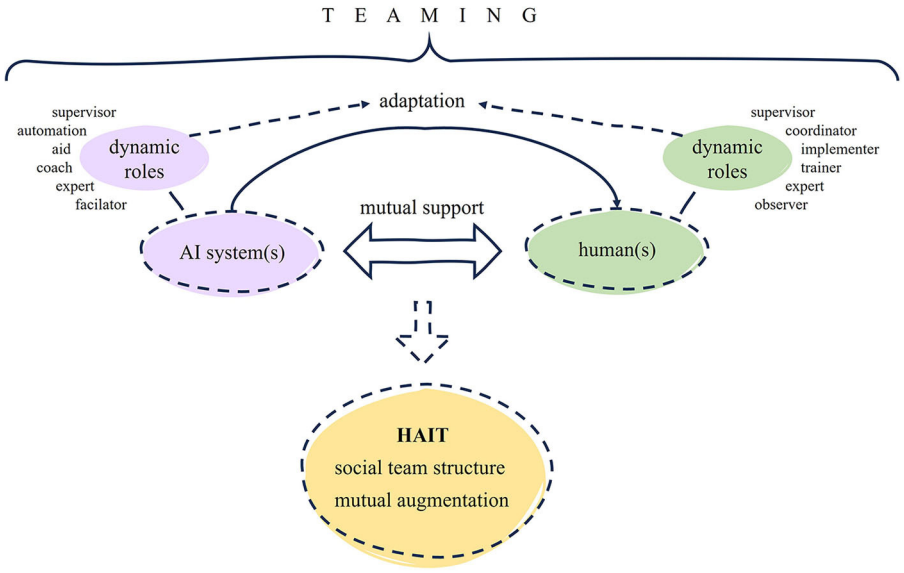

FIGURE 6
The role of AI within HAIT.

systems as genuine team members. They should be capable of, e.g., supporting, taking over, cooperating, or setting borders for the human as needed in specific situation. This view of AI as a dynamic team member, akin to humans, can lead to a new, more profound and nuanced understanding of teaming, which now requires a clear definition and appropriate research efforts.

5.3.3. A definition of human-AI teaming

The need for common ground in HAIT research pointed out throughout this paper as well as the whole Frontiers Research Topic "Human-centered AI at work: Common ground and theories and methods", can, after collating the insights from our review, only be met by a uniting, clear, interdisciplinarity usable definition that is embedded within the idea of socio-technical systems and humane work design. While a diverse research field and evolving insights from different disciplines require the "freedom" to find their own path toward a construct, there comes a point in time where synchronization and integration of perspectives, and necessarily also terminology, become inevitable. This is especially crucial for interdisciplinary exchange, discoverability of publications, discussions employing the same mental models, and transdisciplinary cooperations with practice. Consistent terminology, based on clearly defined and explicit concepts, is a vital prerequisite.

After the field of HAIT research has flourished and produced many valuable insights on various various aspects from different disciplines, the time has come for synchronization. As we could not find an appropriate and integrating definition within our literature search, we decided to use the insights from this review, unite them with the theoretical background in human teaming and develop our own definition of HAIT to answer *RQ4*. We base this definition on (1) the theoretical background presented within this paper of human-machine interaction, (2) the theoretical background on human teaming, especially the skill dimensions by Cannon-Bowers et al. (1995), (3) the terms used within the literature on HAIT, and (4) the definitory elements that the different clusters and papers offered. Figure 7 shows an overview of the definitory aspects that we identified throughout this review, together with exemplary sources.

What we propose is a definition of HAIT that is broad enough to unite different research streams yet focuses on the processes and characteristics of teaming rather that that specific to the technology or the human part. This definition enables joint optimization of social and AI-system(s) as they are both equal parts within it and the focal point is the team as a synergetic socio-technical system:

Human-AI teaming is a process between one or more human(s) and one or more (partially) autonomous AI system(s) acting as team members with unique and complementary capabilities, who work interdependently toward a common goal. The team members' roles are dynamically adapting throughout the collaboration, requiring coordination and mutual communication to meet each other's and the task's requirements. For this, a mutual sharing of intents, shared situational awareness and developing shared mental models are necessary, as well as trust within the team.

Our definition centers on the team level, acknowledging its dynamic and changeable nature by understanding HAIT as a process. This emphasis is a response to the prevailing literature on HAIT, which especially highlights the dynamic and adaptive aspects of teaming (e.g., Hauptman et al., 2023). By understanding teaming as a dynamic process, the collaboration system as a whole becomes more flexible compared to narrowly predefined roles and modes of collaboration. This emphasis stems from the recognition of the diverse capabilities and potential applications of AI systems, which have a significant impact on

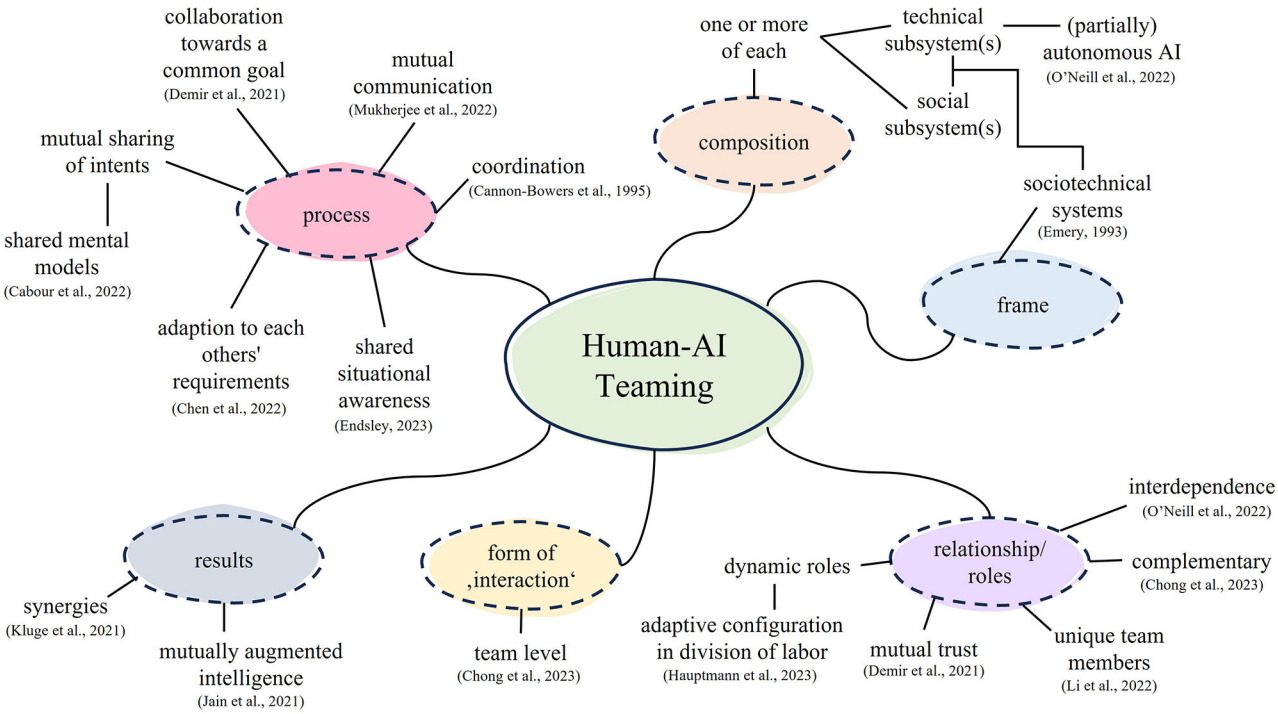

FIGURE 7
Key components of our proposed HAIT definition. The sources indicated in the figure are exemplary.

collaboration modes and possibilities. Moreover, the learning ability of AI systems allows their capabilities to evolve and adapt over time (e.g., Mukherjee et al., 2022), further impacting their potential applications. Emphasizing dynamism and adaptivity enables directly addressing of constantly changing contextual and task-related aspects and requirements. Thus, we consider this aspect crucial in our definition, setting it apart from previous definitions, e.g., by O'Neill et al. (2022).

Nevertheless, we do not perceive our definition as a counter-position to O'Neill et al. (2022). On the contrary, all aspects of their definition can be found within ours, making it an extension offering a different focus, namely on the team process, which we identified as a currently blind spot in the literature. Consequently, we have diverged from including specific capabilities of either subsystem in our definition. We have chosen to focus solely on team-level capabilities that contribute to the success of human-AI teams (e.g., shared situational awareness or shared mental models). This choice acknowledges the potential changes in subsystem capabilities resulting from the dynamics and adaptivity of collaboration.

By centering our definition on team processes and capabilities, we hope to offer a useful definition for future research, building upon current research streams on HAIT and considering insights on human teams.

6. Key takeaways

Navigating through the field of research, the findings from both our network and content analysis and our interpretation of the results, we want to give the five key findings of the review in Figure 8, each of them leading to a specific practical or theoretical implication.

From a practical point of view, we can conclude that human-AI teaming is still in its infancy. Nevertheless, we see great scientific interest in it as well as many antecedents and outcomes that we already have plenty of knowledge on. Practice, from our point of view, should take inspiration from the fast-evolving research and implement human-AI teaming workplaces. Although this takes much more organizational and work redesign and a more creative and generative approach than just to implement AI as a tool, the opportunities are promising for economic reasons as well as humane work.

7. Conclusion

Human-AI teaming is a currently flourishing, multidisciplinary, yet mostly unsystematically approached and so far, one-sided research field. Nevertheless, there is a high need and interest in advancing interdisciplinarity, taking an integrated perspective and finding ways to describe and research a new quality of human collaboration with autonomous technologies, going beyond replacement or mere support of humans in work contexts. Our bibliometric network analysis and scoping review has shown different research streams, understandings, antecedents, and outcomes, revealing the need for a common ground. We close our work by delivering a definition of HAIT considering all the topics from the literature, broadening them with classical teaming knowledge and embedding them in a socio-technical

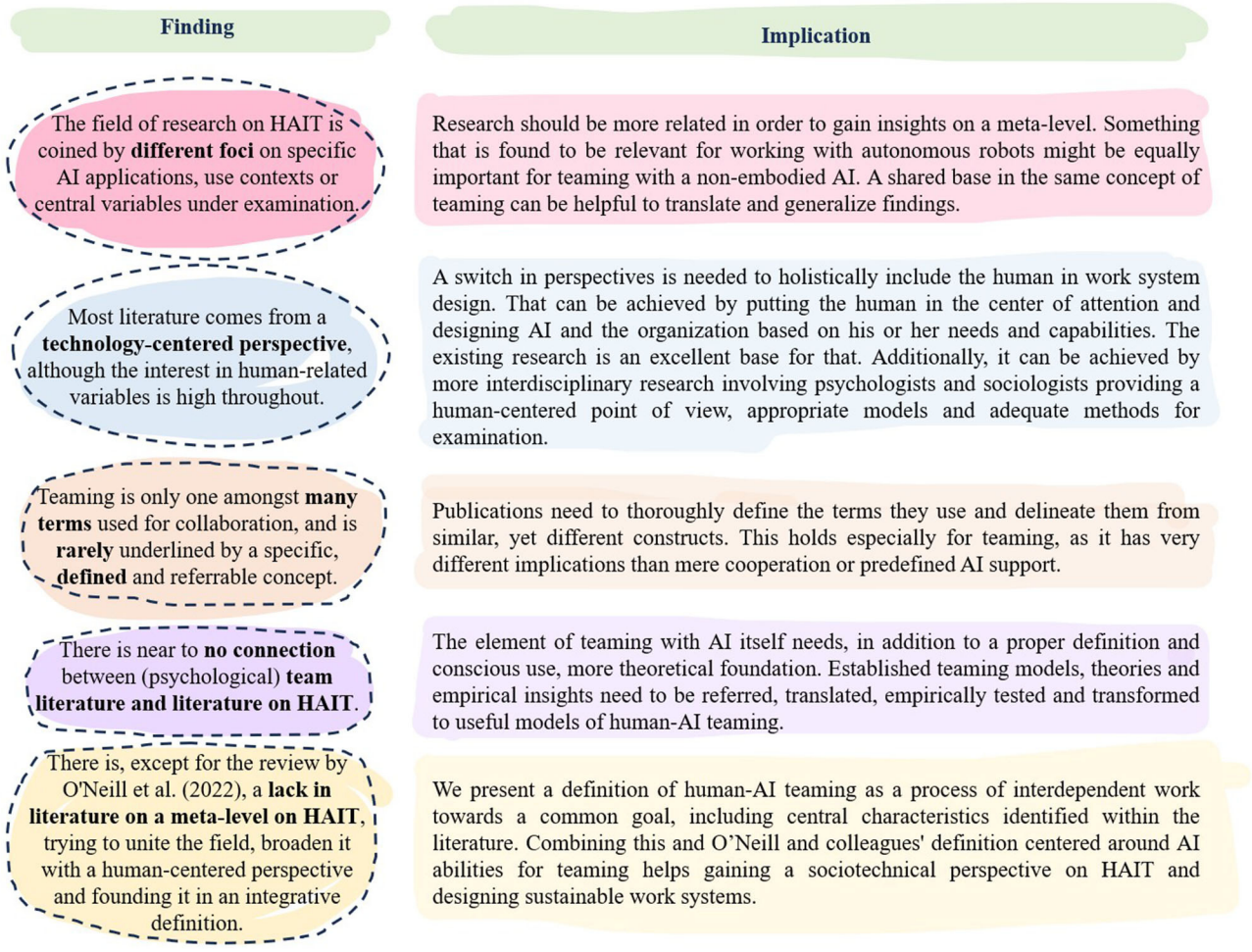

FIGURE 8
Findings and implications of our work.

perspective. By this, we want to stimulate future research and promote the convergence of disparate research streams, ultimately fostering the concept of joint optimization in the context of human-AI teams.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

SB and GO mainly developed the idea of the study. GO took the lead in pre-registration. BG was responsible for the bibliographic network calculation and writing of the respective method sections. AT, GO, and SB were equally responsible for data and content analysis of the articles. SB and AT were in charge of writing the article and of interpreting the results. AK, GO, and BG wrote parts of the paper and, together with CP revised the draft several times to the current state. All authors contributed to the article and approved the submitted version.

Funding

Part of this work was funded by the HUMAINE research project, which was funded by the German Federal Ministry of Education and Research in the program Zukunft der Wertschöpfung—Forschung zu Produktion, Dienstleistung und Arbeit and supervised by Projektträger Karlsruhe (PTKA) (funding code: 02L19C200). In HUMAINE, human-centered work under AI usage is researched, as well as the implementation and realization of human-AI teaming workplaces.

Acknowledgments

We acknowledge support by the Open Access Publication Funds of the Ruhr-Universität Bochum.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2023.1250725/full#supplementary-material

References

Aliev, K., and Antonelli, D. (2021). Proposal of a monitoring system for collaborative robots to predict outages and to assess reliability factors exploiting machine learning. *Appl. Sci.* 11, 1–20. doi: 10.3390/app11041621

Appelbaum, S. H. (1997). Socio-technical systems theory: an intervention strategy for organizational development. *Manag. Decis.* 35, 452–463. doi: 10.1108/00251749710173823

Aria, M., and Cuccurullo, C. (2017). Bibliometrix: an R-tool for comprehensive science mapping analysis. *J. Informetr.* 11, 959–975. doi: 10.1016/j.joi.2017.08.007

Arksey, H., and O'Malley, L. (2005). Scoping studies: towards a methodological framework. *Int. J. Soc. Res. Methodol.* 8, 19–32. doi: 10.1080/1364557032000119616

Armstrong, R., Hall, B. J., Doyle, J., and Waters, E. (2011). Cochrane update. 'Scoping the scope' of a cochrane review. *J. Public Health* 33, 147–150. doi: 10.1093/pubmed/fdr015

Arslan, A., Cooper, C., Khan, Z., Golgeci, I., and Ali, I. (2022). Artificial intelligence and human workers interaction at team level: a conceptual assessment of the challenges and potential HRM strategies. *Int. J. Manpow.* 43, 75–88. doi: 10.1108/IJM-01-2021-0052

Aversa, P., Cabantous, L., and Haefliger, S. (2018). When decision support systems fail: insights for strategic information systems from Formula 1. *J. Strat. Inf. Syst.* 27, 221–236. doi: 10.1016/j.jsis.2018.03.002

Berretta, S., Tausch, A., Peifer, C., and Kluge, A. (2023). The Job Perception Inventory: considering human factors and needs in the design of human–AI work. *Front. Psychol.* 14, 1128945. doi: 10.3389/fpsyg.2023.1128945

Bornmann, L., and Daniel, H.-D. (2008). What do citation counts measure? A review of studies on citing behavior. *J. Document.* 64, 45–80. doi: 10.1108/00220410810844150

Boyack, K. W., and Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *J. Am. Soc. Inf. Sci. Technol.* 61, 2389–2404. doi: 10.1002/asi.21419

Bredahl, L. (2022). Bibliometric tools for academia. *Libr. Technol. Rep.* 8, 12–21. doi: 10.5860/ltr.58n8

Breque, M., de Nul, L., and Petridis, A. (2021). *Industry 5.0: Towards a Sustainable, Human-Centric and Resilient European industry. R and I Paper Series, Policy Brief.* Brussels: Publications Office of the European Union.

Bütepage, J., and Kragic, D. (2017). Human-robot collaboration: from psychology to social robotics. *arXiv preprint* arXiv:1705.10146. doi: 10.48550/arXiv.1705.10146

Cabitza, F., Campagner, A., and Simone, C. (2021). The need to move away from agential-AI: empirical investigations, useful concepts and open issues. *Int. J. Hum. Comput. Stud.* 155, 102696. doi: 10.1016/j.ijhcs.2021.102696

Cabour, G., Morales-Forero, A., Ledoux, É., and Bassetto, S. (2022). An explanation space to align user studies with the technical development of explainable AI. *AI Soc.* 38, 869–887. doi: 10.1007/s00146-022-01536-6

Cannon-Bowers, J. A., Tannenbaum, S. I., Salas, E., and Volpe, C. E. (1995). "Defining team competencies and establishing team training requirements," in *Team Effectiveness and Decision Making in Organizations*, eds R. Guzzo, and E. Salas (San Francisco, CA: Jossey-Bass), 333–380.

Casey-Campbell, M., and Martens, M. L. (2009). Sticking it all together: a critical assessment of the group cohesion-performance literature. *Int. J. Manag. Rev.* 11, 223–246. doi: 10.1111/j.1468-2370.2008.00239.x

Castro, A., Silva, F., and Santos, V. (2021). Trends of human-robot collaboration in industry contexts: handover, learning, and metrics. *Sensors* 21. doi: 10.3390/s21124113

Chandel, A., and Sharma, B. (2023). "Technology aspects of artificial intelligence: industry 5.0 for organization decision making," in *Lecture Notes in Networks and Systems: Vol. 521, Information Systems and Management Science: Conference Proceedings of 4th International Conference on Information Systems and Management Science (ISMS) 2021*, eds L. Garg, D. S. Sisodia, N. Kesswani, J. G. Vella, I. Brigui, P. Xuereb, et al. (Cham: Springer International Publishing), 79–90.

Chazette, L., Brunotte, W., and Speith, T. (2021). "Exploring explainability: a definition, a model, and a knowledge catalogue," in *2021 IEEE 29th International Requirements Engineering Conference (RE)* (Notre Dame, IN: IEEE).

Chen, Q. Z., Schnabel, T., Nushi, B., and Amershi, S. (2022). "Hint: integration testing for AI-based features with humans in the loop," in *ACM Digital Library, Proceedings of the 27th International Conference on Intelligent User Interfaces* (Helsinki: Association for Computing Machinery), 549–565.

Chong, L., Raina, A., Goucher-Lambert, K., Kotovsky, K., and Cagan, J. (2023). The evolution and impact of human confidence in artificial intelligence and in themselves on AI-assisted decision-making in design. *J. Mech. Des.* 145, 031401. doi: 10.1115/1.4055123

Cruz, F., Dazeley, R., Vamplew, P., and Moreira, I. (2021). Explainable robotic systems: understanding goal-driven actions in a reinforcement learning scenario. *Neural Comp. Appl.* doi: 10.1007/s00521-021-06425-5. [Epub ahead of print].

Csárdi, G., Nepusz, T., Müller, K., Horvát, S., Traag, V., Zanini, F., et al. (2023). *igraph for R: R Interface of the igraph Library for Graph Theory and Network Analysis.* Zenodo. Available online at: https://CRAN.R-project.org/package$=$igraph (accessed March 30, 2023).

Cuevas, H. M., Fiore, S. M., Caldwell, B. S., and Strater, L. (2007). Augmenting team cognition in human-automation teams performing in complex operational environments. *Aviat. Space Environ. Med.* 78(5Suppl.), B63B70.

Dahl, M., Larsen, C., Eros, E., Bengtsson, K., Fabian, M., and Falkman, P. (2022). Interactive formal specification for efficient preparation of intelligent automation systems. *CIRP J. Manufact. Sci. Technol.* 38, 129–138. doi: 10.1016/j.cirpj.2022.04.013

Dellermann, D., Calma, A., Lipusch, N., Weber, T., Weigel, S., and Ebel, P. (2021). *The Future of Human-AI Collaboration: a Taxonomy of Design Knowledge for Hybrid Intelligence Systems.*

Demir, M., McNeese, N. J., Gorman, J. C., Cooke, N. J., Myers, C. W., and Grimm, D. A. (2021). Exploration of teammate trust and interaction dynamics in human-autonomy teaming. *IEEE Transact. Hum. Mach. Syst.* 51, 696–705. doi: 10.1109/THMS.2021.3115058

Deutsch, M. (1949). A theory of co-operation and competition. *Hum. Relat.* 2, 129–152. doi: 10.1177/001872674900200204

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., and Lim, W. M. (2021). How to conduct a bibliometric analysis: an overview and guidelines. *J. Bus. Res.* 133, 285–296. doi: 10.1016/j.jbusres.2021.04.070

Donthu, N., Kumar, S., and Pattnaik, D. (2020). Forty-five years of journal of business research: a bibliometric analysis. *J. Bus. Res.* 109, 1–14. doi: 10.1016/j.jbusres.2019.10.039

Dubey, A., Abhinav, K., Jain, S., Arora, V., and Puttaveerana, A. (2020). "HACO: a framework for developing human-AI teaming," in *Proceedings of the 13th Innovations in Software Engineering Conference on Formerly Known as India Software Engineering Conference*, eds S. Jain, A. Gupta, D. Lo, D. Saha, and R. Sharma (ACM), 1–9. doi: 10.1145/3385032.3385044

Ebadi, A., and Schiffauerova, A. (2015). On the relation between the small world structure and scientific activities. *PLoS ONE* 10, e0121129. doi: 10.1371/journal.pone.0121129

Emery, F. (1993). "Characteristics of socio-technical systems," in *The Social Engagement of Social Science, Vol. 2*, eds E. Trist, H. Murray, and B. Trist (Pennsylvania: University of Pennsylvania Press).

Endsley, M. R. (2023). Supporting human-AI teams: transparency, explainability, and situation awareness. *Comput. Human Behav.* 140, 107574. doi: 10.1016/j.chb.2022.107574

Fan, M., Yang, X., Yu, T., Liao, Q. V., and Zhao, J. (2022). Human-AI collaboration for UX evaluation: effects of explanation and synchronization. *Proc. ACM Hum. Comp. Interact.* 6, 96. doi: 10.1145/3512943

Flake, J. K., and Fried, E. I. (2020). Measurement Schmeasurement: questionable measurement practices and how to avoid them. *Adv. Methods Pract. Psychol. Sci.* 3, 456–465. doi: 10.1177/2515245920952393

Fogliato, R., Chappidi, S., Lungren, M., Fisher, P., Wilson, D., Fitzke, M., et al. (2022). "Who goes first? Influences of human-AI workflow on decision making in clinical imaging," in *ACM Digital Library, Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY: Association for Computing Machinery), 1362–1374.

Galin, R., and Meshcheryakov, R. (2021). "Collaborative robots: development of robotic perception system, safety issues, and integration of AI to imitate human behavior," in *Smart Innovation, Systems and Technologies. Proceedings of 15th International Conference on Electromechanics and Robotics "Zavalishin's Readings"*, Vol. 187, eds A. Ronzhin, and V. Shishlakov (Singapore: Springer), 175–185.

Gupta, P., and Woolley, A. W. (2021). "Articulating the role of artificial intelligence in collective intelligence: a transactive systems framework," in *65th Human Factors and Ergonomics Society Annual Meeting (HFES 2021)* (Maryland: Curran Associates Inc.), 670–674.

Hauptman, A. I., Schelble, B. G., McNeese, N. J., and Madathil, K. C. (2023). Adapt and overcome: perceptions of adaptive autonomous agents for human-AI teaming. *Comput. Human Behav.* 138, 107451. doi: 10.1016/j.chb.2022.107451

Hevey, D. (2018). Network analysis: a brief overview and tutorial. *Health Psychol. Behav. Med.* 6, 301–328. doi: 10.1080/21642850.2018.1521283

Huang, B., Huan, Y., Da Xu, L., Zheng, L., and Zou, Z. (2019). Automated trading systems statistical and machine learning methods and hardware implementation: a survey. *Enterprise Inf. Syst.* 13, 132–144. doi: 10.1080/17517575.2018.1493145

Huchler, N. (2015). Die "Rolle des Menschen" in der Industrie 4.0 - Technikzentrierter vs. humanzentrierter Ansatz. *AIS Stud.* 9, 57–79. doi: 10.21241/SSOAR.64826

Hughes, C., Robert, L., Frady, K., and Arroyos, A. (2019). "Artificial intelligence, employee engagement, fairness, and job outcomes," in *Managing Technology and Middle- and Low-skilled Employees*, eds C. Hughes, L. Robert, K. Frady, and A. Arroyos (Bingley: Emerald Publishing Limited), 61–68.

Jain, H., Padmanabhan, B., Pavlou, P. A., and Raghu, T. S. (2021). Editorial for the special section on humans, algorithms, and augmented intelligence: the future of work, organizations, and society. *Inf. Syst. Res.* 32, 675–687. doi: 10.1287/isre.2021.1046

Jain, R., Garg, N., and Khera, S. N. (2022). Effective human–AI work design for collaborative decision-making. *Kybernetes*. doi: 10.1108/K-04-2022-0548. [Epub ahead of print].

Jarneving, B. (2005). A comparison of two bibliometric methods for mapping of the research front. *Scientometrics* 65, 245–263. doi: 10.1007/s11192-005-0270-7

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: human-AI symbiosis in organizational decision making. *Bus. Horiz.* 61, 577–586. doi: 10.1016/j.bushor.2018.03.007

Jiang, N., Liu, X., Liu, H., Lim, E. T. K., Tan, C.-W., and Gu, J. (2022). Beyond AI-powered context-aware services: the role of human–AI collaboration. *Ind. Manag. Data Syst.* doi: 10.1108/IMDS-03-2022-0152. [Epub ahead of print].

Johnson, C. J., Demir, M., McNeese, N. J., Gorman, J. C., Wolff, A. T., and Cooke, N. J. (2021). The impact of training on human-autonomy team communications and trust calibration. *Hum. Fact.* 187208211047323. doi: 10.1177/00187208211047323

Kauffeld, S. (2001). *Teamdiagnose [Team Diagnosis]*. Göttingen: Verlag für Angewandte Psychologie.

Kho, M. E., and Brouwers, M. C. (2012). The systematic review and bibliometric network analysis (SeBriNA) is a new method to contextualize evidence. Part 1: description. *J. Clin. Epidemiol.* 65, 1010–1015. doi: 10.1016/j.jclinepi.2012.03.009

Kleminski, R., Kazienko, P., and Kajdanowicz, T. (2022). Analysis of direct citation, co-citation and bibliographic coupling in scientific topic identification. *J. Inf. Sci.* 48, 349–373. doi: 10.1177/0165551520962775

Kluge, A., Ontrup, G., Langholf, V., and Wilkens, U. (2021). Mensch-KI-Teaming: Mensch und Künstliche Intelligenz in der Arbeitswelt von morgen. *ZWF Zeitschrift fuer Wirtschaftlichen Fabrikbetrieb* 116, 728–734. doi: 10.1515/zwf-2021-0112

Kozlowski, S. W. J., and Bell, B. S. (2012). "Work groups and teams in organizations: review update," in *Handbook of Psychology, Industrial and Organizational Psychology: Industrial and Organizational Psychology, 2nd Edn.*, eds I. Weiner, N. W. Schmitt, and S. Highhouse (Chichester: Wiley).

Kraus, M., Wagner, N., and Minker, W. (2021). "Modelling and predicting trust for developing proactive dialogue strategies in mixed-initiative interaction," in (*ACM Digital Library. Proceedings of the 2021 International Conference on Multimodal Interaction*, ed Z. Hammal (Association for Computing Machinery), 131–140.

Kridalukmana, R., Lu, H., and Naderpour, M. (2022). Self-explaining abilities of an intelligent agent for transparency in a collaborative driving context. *IEEE Transact. Hum. Mach. Syst.* 52, 1155–1165. doi: 10.1109/THMS.2022.3202900

Kumpulainen, M., and Seppänen, M. (2022). Combining Web of Science and Scopus datasets in citation-based literature study. *Scientometrics* 127, 5613–5631. doi: 10.1007/s11192-022-04475-7

Kusters, R., Misevic, D., Berry, H., Cully, A., Le Cunff, Y., Dandoy, L., et al. (2020). Interdisciplinary research in artificial intelligence: challenges and opportunities. *Front. Big Data* 3, 577974. doi: 10.3389/fdata.2020.577974

Lai, V., Carton, S., Bhatnagar, R., Liao, Q. V., Zhang, Y., and Tan, C. (2022). "Human-AI collaboration via conditional delegation: a case study of content moderation," in *ACM Digital Library, Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA), ed S. Barbosa (New York, NY: Association for Computing Machinery).

Le, K. B. Q., Sajtos, L., and Fernandez, K. V. (2023). Employee-(ro)bot collaboration in service: an interdependence perspective. *J. Serv. Manag.* 34, 176–207. doi: 10.1108/JOSM-06-2021-0232

Lee, J. D., and See, K. A. (2004). Trust in automation: designing for appropriate reliance. *Hum. Fact.* 46, 50–80. doi: 10.1518/hfes.46.1.50.30392

Li, J., Huang, J., Liu, J., and Zheng, T. (2022). Human-AI cooperation: modes and their effects on attitudes. *Telemat. Informat.* 73, 101862. doi: 10.1016/j.tele.2022.101862

Lima, S., and de Assis Carlos Filho, F. (2019). Bibliometric analysis of scientific production on sharing economy. *Revista De Gestão* 26, 237–255. doi: 10.1108/REGE-01-2019-0018

Lyons, J. B., Sycara, K., Lewis, M., and Capiola, A. (2021). Human-autonomy teaming: definitions, debates, and directions. *Front. Psychol.* 12, 589585. doi: 10.3389/fpsyg.2021.589585

Mathieu, J. E., Hollenbeck, J. R., van Knippenberg, D., and Ilgen, D. R. (2017). A century of work teams in the Journal of Applied Psychology. *J. Appl. Psychol.* 102, 452–467. doi: 10.1037/apl0000128

McNeese, N. J., Demir, M., Cooke, N. J., and Myers, C. (2018). Teaming with a synthetic teammate: insights into human-autonomy teaming. *Hum. Fact.* 60, 262–273. doi: 10.1177/0018720817743223

McNeese, N. J., Demir, M., Cooke, N. J., and She, M. (2021). Team situation awareness and conflict: a study of human–machine teaming. *J. Cognit. Eng. Decis. Making* 15, 83–96. doi: 10.1177/15553434211017354

Moher, D., Liberati, A., Tetzlaff, J., and Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med.* 6, e1000097. doi: 10.1371/journal.pmed.1000097

Mukherjee, D., Gupta, K., Chang, L. H., and Najjaran, H. (2022). A survey of robot learning strategies for human-robot collaboration in industrial settings. *Robot. Comput. Integr. Manuf.* 73, 102231. doi: 10.1016/j.rcim.2021.102231

Munn, Z., Peters, M. D. J., Stern, C., Tufanaru, C., McArthur, A., and Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Med. Res. Methodol.* 18, 143. doi: 10.1186/s12874-018-0611-x

Naiseh, M., Al-Thani, D., Jiang, N., and Ali, R. (2023). How the different explanation classes impact trust calibration: the case of clinical decision support systems. *Int. J. Hum. Comput. Stud.* 169. doi: 10.1016/j.ijhcs.2022.102941

Nass, C., Fogg, B. J., and Moon, Y. (1996). Can computers be teammates? *Int. J. Hum. Comput. Stud.* 45, 669–678. doi: 10.1006/ijhc.1996.0073

National Academies of Sciences, Engineering, and Medicine (2021). *Human-AI Teaming: State of the Art and Research Needs*. Washington, DC: The National Academies Press.

Navarro, J., Heuveline, L., Avril, E., and Cegarra, J. (2018). Influence of human-machine interactions and task demand on automation selection and use. *Ergonomics* 61, 1601–1612. doi: 10.1080/00140139.2018.1501517

Newman, M. E. J., and Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E* 69, 26113. doi: 10.1103/PhysRevE.69.026113

O'Neill, T., McNeese, N. J., Barron, A., and Schelble, B. G. (2022). Human-autonomy teaming: a review and analysis of the empirical literature. *Hum. Fact.* 64, 904–938. doi: 10.1177/0018720820960865

Othman, U., and Yang, E. (2022). "An overview of human-robot collaboration in smart manufacturing," in *2022 27th International Conference on Automation and Computing (ICAC)* (Bristol: IEEE), 1–6.

Ouzzani, M., Hammady, H., Fedorowicz, Z., and Elmagarmid, A. (2016). Rayyan-a web and mobile app for systematic reviews. *Syst. Rev.* 5, 210. doi: 10.1186/s13643-016-0384-4

Parasuraman, R., Sheridan, T. B., and Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transact. Syst. Man Cybernet.* 30, 286–297. doi: 10.1109/3468.844354

Parker, S. K., Morgeson, F. P., and Johns, G. (2017). One hundred years of work design research: looking back and looking forward. *J. Appl. Psychol.* 102, 403–420. doi: 10.1037/apl0000106

Picon, A. (2004). Engineers and engineering history: problems and perspectives. *Hist. Technol.* 20, 421–436. doi: 10.1080/0734151042000304367

Pynadath, D. V., Gurney, N., and Wang, N. (2022). "Explainable reinforcement learning in human-robot teams: the impact of decision-tree explanations on transparency," in *Proceedings of the 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (Napoli: IEEE), 749–756.

R Core Team (2022). *R: A Language and Environment for Statistical Computing [Computer software]*. Vienna: R Foundation for Statistical Computing. Available online at: https://www.R-project.org/ (accessed March 30, 2023).

Rastogi, C., Zhang, Y., Wei, D., Varshney, K. R., Dhurandhar, A., and Tomsett, R. (2022). Deciding fast and slow: the role of cognitive biases in AI-assisted decision-making. *Proc. ACM Hum. Comp. Interact.* 6, 1–22. doi: 10.1145/3512930

Rix, J. (2022). "From tools to teammates: conceptualizing humans' perception of machines as teammates with a systematic literature review," in *55th Hawaii International Conference on System Sciences (HICSS)* (Maui, HI).

Rodrigues, I. R., Dantas, M., de Oliveira Filho, A. T., Barbosa, G., Bezerra, D., Souza, R. S., et al. (2023). A framework for robotic arm pose estimation and movement prediction based on deep and extreme learning models. *J. Supercomput.* 79, 7176–7205. doi: 10.1007/s11227-022-04936-z

Roethlisberger, F. J., and Dickson, W. J. (1939). *Management and the Worker: An Account of Research Program Conducted by the Western Electric Company, Hawthorne Works*. Cambridge, MA: Harvard University Press.

Salas, E., Burke, C. S., and Cannon-Bowers, J. A. (2000). Teamwork: emerging principles. *Int. J. Manag. Rev.* 2, 339–356. doi: 10.1111/1468-2370.00046

Saßmannshausen, T., Burggräf, P., Wagner, J., Hassenzahl, M., Heupel, T., and Steinberg, F. (2021). Trust in artificial intelligence within production management - an exploration of antecedents. *Ergonomics* 64, 1333–1350. doi: 10.1080/00140139.2021.1909755

Schmidtler, J., Knott, V., Hölzel, C., and Bengler, K. (2015). Human centered assistance applications for the working environment of the future. *Occup. Ergon.* 12, 83–95. doi: 10.3233/OER-150226

Seeber, I., Waizenegger, L., Seidel, S., Morana, S., Benbasat, I., and Lowry, P. B. (2020). Collaborating with technology-based autonomous agents. *Int. Res.* 30, 1–18. doi: 10.1108/INTR-12-2019-0503

Semeraro, F., Griffiths, A., and Cangelosi, A. (2022). Human–robot collaboration and machine learning: a systematic review of recent research. *Robot. Comput. Integr. Manuf.* 79. doi: 10.1016/j.rcim.2022.102432

Silva, A., Schrum, M., Hedlund-Botti, E., Gopalan, N., and Gombolay, M. (2022). Explainable artificial intelligence: evaluating the objective and subjective impacts of xAI on human-agent interaction. *Int. J. Hum. Comp. Interact.* 1390–1404. doi: 10.1080/10447318.2022.2101698

Šukjurovs, I., Zvirgzdina, R., and Jeromanova-Maura, S. (2019). Artificial intelligence in workplaces and how it will affect employment in latvia. *Environ. Technol. Resour.* 2, 154. doi: 10.17770/etr2019vol2.4151

Sundstrom, E., McIntyre, M., Halfhill, T., and Richards, H. (2000). Work groups: from the Hawthorne studies to work teams of the 1990s and beyond. *Group Dyn.* 4, 44–67. doi: 10.1037/1089-2699.4.1.44

Tabrez, A., Luebbers, M. B., and Hayes, B. (2022). "Descriptive and prescriptive visual guidance to improve shared situational awareness in human-robot teaming," in *ACM Digital Library, Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, ed C. Pelachaud (International Foundation for Autonomous Agents and Multiagent Systems), 1256–1264.

Thibaut, J. W., and Kelley, H. H. (1959). *The Social Psychology of Groups*. New York, NY: Wiley.

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., et al. (2018). Prisma extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Ann. Intern. Med.* 169, 467–473. doi: 10.7326/M18-0850

Tsai, C.-Y., Marshall, J. D., Choudhury, A., Serban, A., Tsung-Yu Hou, Y., Jung, M. F., et al. (2022). Human-robot collaboration: a multilevel and integrated leadership framework. *Leadersh. Q.* 33, 101594. doi: 10.1016/j.leaqua.2021.101594

Vecchio, R. P., and Appelbaum, S. H. (1995). *Managing organizational behaviour: A Canadian perspective. Dryden Series in Management*. Toronto, ON: Dryden.

Verhagen, R. S., Neerincx, M. A., and Tielman, M. L. (2022). The influence of interdependence and a transparent or explainable communication style on human-robot teamwork. *Front. Robot. AI* 9, 993997. doi: 10.3389/frobt.2022.993997

Vössing, M., Kühl, N., Lind, M., and Satzger, G. (2022). Designing transparency for effective human-AI collaboration. *Inf. Syst. Front.* 24, 877–895. doi: 10.1007/s10796-022-10284-3

Walliser, J. C., de Visser, E. J., Wiese, E., and Shaw, T. H. (2019). Team structure and team building improve human–machine teaming with autonomous agents. *J. Cognit. Eng. Decis. Making* 13, 258–278. doi: 10.1177/1555343419867563

Wang, D., Maes, P., Ren, X., Shneiderman, B., Shi, Y., and Wang, Q. (2021). "Designing AI to work WITH or FOR people?," in *ACM Digital Library. Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, ed Y. Kitamura (New York, NY: Association for Computing Machinery).

Weisz, J. D., Muller, M., Houde, S., Richards, J., Ross, S. I., Martinez, F., et al. (2021). "Perfection not required? Human-AI partnerships in code translation," in *Proceedings of the 26th International Conference on Intelligent User Interfaces*, eds T. Hammond, K. Verbert, D. Parra, B. Knijnenburg, J. O'Donovan, and P. Teale (New York, NY: ACM), 402–412.

Wilkens, U., Langholf, V., Ontrup, G., and Kluge, A. (2021). "Towards a maturity model of human-centered AI - A reference for AI implementation at the workplace," in *Competence Development and Learning Assistance Systems for the Data-Driven Future*, eds W. Sihn and S. Schlund (Goto: Verlag), 179–198.

Xiong, W., Wang, C., and Ma, L. (2023). Partner or subordinate? Sequential risky decision-making behaviors under human-machine collaboration contexts. *Comp. Hum. Behav.* 139, 107556. doi: 10.1016/j.chb.2022.107556

Yam, K. C., Tang, P. M., Jackson, J. C., Su, R., and Gray, K. (2023). The rise of robots increases job insecurity and maladaptive workplace behaviors: multimethod evidence. *J. Appl. Psychol.* 108, 850–870. doi: 10.1037/apl0001045

Yang, N. Y., and Choi, J. S. (2014). Relationships of nurses' perception, nursing performance, job stress, and burnout in relation to the joint commission international hospital accreditation. *J. Kor. Acad. Nurs. Administr.* 20, 1. doi: 10.11111/jkana.2014.20.1.1

Yang, Z., Algesheimer, R., and Tessone, C. J. (2016). A comparative analysis of community detection algorithms on artificial networks. *Sci. Rep.* 6, 30750. doi: 10.1038/srep30750

Zhang, L., and Amos, C. (2023). Dignity and use of algorithm in performance evaluation. *Behavi. Inf. Technol.* 1–18. doi: 10.1080/0144929X.2022.2164214

frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 27 October 2023
DOI 10.3389/frai.2023.1257057

OPEN ACCESS

EDITED BY
Annette Kluge,
Ruhr University Bochum, Germany

REVIEWED BY
Benedikt Leichtmann,
Johannes Kepler University of Linz, Austria
Marcus Grum,
University of Potsdam, Germany

*CORRESPONDENCE
Athanasios Mazarakis
✉ ama@informatik.uni-kiel.de
Christian Bernhard-Skala
✉ bernhard-skala@die-bonn.de
Martin Braun
✉ martin.braun@iao.fraunhofer.de
Isabella Peters
✉ ipe@informatik.uni-kiel.de

RECEIVED 11 July 2023
ACCEPTED 05 October 2023
PUBLISHED 27 October 2023

CITATION
Mazarakis A, Bernhard-Skala C, Braun M and Peters I (2023) What is critical for human-centered AI at work? – Toward an interdisciplinary theory.
Front. Artif. Intell. 6:1257057.
doi: 10.3389/frai.2023.1257057

What is critical for human-centered AI at work? – Toward an interdisciplinary theory

Athanasios Mazarakis[1]*, Christian Bernhard-Skala[2]*, Martin Braun[3]* and Isabella Peters[1]*

[1]ZBW – Leibniz Information Centre for Economics, Web Science, Kiel, Germany, [2]Department of Organisation and Program Planning, German Institute for Adult Education - Leibniz Centre for Lifelong Learning, Bonn, Germany, [3]Fraunhofer Institute for Industrial Engineering, User Experience, Stuttgart, Germany

Human-centered artificial intelligence (HCAI) has gained momentum in the scientific discourse but still lacks clarity. In particular, disciplinary differences regarding the scope of HCAI have become apparent and were criticized, calling for a systematic mapping of conceptualizations—especially with regard to the work context. This article compares how human factors and ergonomics (HFE), psychology, human-computer interaction (HCI), information science, and adult education view HCAI and discusses their normative, theoretical, and methodological approaches toward HCAI, as well as the implications for research and practice. It will be argued that an interdisciplinary approach is critical for developing, transferring, and implementing HCAI at work. Additionally, it will be shown that the presented disciplines are well-suited for conceptualizing HCAI and bringing it into practice since they are united in one aspect: they all place the human being in the center of their theory and research. Many critical aspects for successful HCAI, as well as minimum fields of action, were further identified, such as human capability and controllability (HFE perspective), autonomy and trust (psychology and HCI perspective), learning and teaching designs across target groups (adult education perspective), as much as information behavior and information literacy (information science perspective). As such, the article lays the ground for a theory of human-centered interdisciplinary AI, i.e., the Synergistic Human-AI Symbiosis Theory (SHAST), whose conceptual framework and founding pillars will be introduced.

KEYWORDS

artificial intelligence, interdisciplinary, HCAI, human factors and ergonomics, information science, human-computer interaction (HCI), adult education, psychology

1. Introduction

The excitement around artificial intelligence (AI) is sparking a flurry of activity by researchers, developers, business leaders, and policy-makers worldwide. The promise of groundbreaking advances from machine learning and other algorithms drives discussions and attracts huge investments in, e.g., medical, manufacturing, and military innovations (Shneiderman, 2022). However, much of the debate in society is associated with aspects of whether or not AI will replace people in business activities (Del Giudice et al., 2023). In addition, trust in AI systems, transparency, and explaining such systems is not straightforward to end users (Laato et al., 2022).

In a survey in Germany in May 2023, 46% of 1,220 respondents considered AI technologies to be more of a risk for them personally, while only 39% saw opportunities in AI solutions. However, openness to these new technologies decreases with age and increases

with education: most younger people up to age 34 see AI as an opportunity, as do individuals with a university degree (Infratest dimap, 2023). The results of previous surveys conducted worldwide are confirmed, although the results differ greatly depending on the economic development in each country (Ipsos, 2022).

More particularly, in the workplace, there is a risk of creating a defeatist mentality among the employees when ignoring human aspects of AI implementation. Similar examples exist from the past, e.g., knowledge management faced the same challenges around individual, organizational, and technological barriers (Riege, 2005). Nevertheless, it is argued that high levels of human control and automation are likely to simultaneously empower people and not just emulate humans (Shneiderman, 2020). The idea of human centeredness in AI implementation binds these critical research results together.

Artificial intelligence is, by definition, a sequence of mathematical models created by humans, which are executed by computers. The OECD defines AI more precisely: "An AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. [...] AI systems are designed to operate with varying levels of autonomy" (OECD., 2019, p. 23–24). This is a refinement of the definition of McCarthy (2007, p. 2), in which AI is defined as "[...] the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable."

In general, Artificial Narrow Intelligence (ANI) and Artificial General Intelligence (AGI) are distinguished. An AGI system would be an autonomous agent that can learn unsupervised (McLean et al., 2021). ANI has achieved enormous success in determined situations with a low-dimensional phase space, such as strategy games (Lenzen, 2019). However, the methodology of ANI performance can only be applied to a limited range of tasks (Landgrebe and Smith, 2022).

The organization of working processes may profit from a design approach that integrates human and technical intelligence. Such an organization is achieved, among other things, when humans and machines can use their specific skills and when humans and machines mutually support each other in gaining capabilities (Braun, 2017). In this sense, human-centered AI (HCAI) already consists of a set of standards, concepts, and principles, like e.g., fairness, accountability, beneficence, justice, and explicability, to name a few (Huchler et al., 2020). However, these principles are not consistently implemented in practice because of competing goals of productivity and cost-cutting, which has been a traditional challenge of HFE (Spitzley, 1980).

The present article aims at setting the foundations for an interdisciplinary theory of human-centered AI. For this, the article reflects in Section 6.2 on learnings from an interdisciplinary research project, which is considered as a demonstrator. This demonstrator combines normative, theoretical, and methodological concepts from human factors and ergonomics (HFE), psychology, human-computer interaction (HCI), information science, and adult education to study AI at work.

The article introduces each discipline's perspective on human-centricity and AI at work. It discusses the implications for developing, transferring, and implementing AI at the workplace, including disciplines not yet in the spotlight about HCAI at work. By explaining the fruitful interplay of these disciplines and how they contribute to human-centricity, we will ultimately argue that a theory of human-centered AI will immensely benefit from incorporating critical concepts, norms, and theories from the presented disciplines as well as from an interdisciplinary approach. This article, therefore, sheds light on different co-existing perspectives on and criteria of human-centered AI and will show how they can be meaningfully integrated to answer the research question of *what is critical for human-centered AI at work*. So far, this has only been conducted for, e.g., social sciences (Miller, 2019) and thus has not acknowledged or even included more diverse disciplines. However, many authors have identified an urgent need for collaboration across disciplines for human stakeholders, e.g., for explainable AI (Langer et al., 2021). This article contributes to this research gap by introducing views from heterogeneous disciplines with either a focus on individuals, such as psychology, or a focus on technology, such as HCI.

Moreover, the article leverages insights from disciplines that are either primarily concerned with a work context, such as HFE, or disciplines that study the work context as one research object amongst others, such as adult learning and information science. Furthermore, it systematizes results from the different disciplines by using the five perspectives on human-centricity by Wilkens et al. (2021). In addition, it identifies fields of action for human-centered AI implementation—such as supporting balanced workload, information literacy, providing tailored learning opportunities for low-skilled workers, enhancing technology acceptance, and building trust. By that, it finally sets the foundation for a synergistic human-AI symbiosis theory, which includes these aspects of AI implementation. Overall, this article is written for an interdisciplinary readership interested in human-centered AI. As a result, this article has more of an explorative, descriptive, and conceptual character.

2. Supporting balanced workload: HCAI in human factors and ergonomics

Human factors and ergonomics (HFE) has developed concepts and principles for work design, especially for human-technology interaction. These concepts can be applied to AI systems in terms of a division of functions between humans and AI. The principles of HFE inhibit a mutually reinforcing relationship between humans and AI based on capabilities and ethical design principles.

2.1. Defining ergonomics and illustrating the methodology

Ergonomics is the scientific discipline concerned with the understanding of interactions among humans and other elements of a system; ergonomics is also the profession that applies theory,

principles, data, and methods to design in order to optimize human wellbeing and overall system performance (IEA, 2000). The terms *ergonomics* and *human factors* are often used interchangeably or as a unit (HFE). Domains of HFE are physical ergonomics (i.e., human anthropometric, physiological, and biomechanical characteristics as they relate to work), cognitive ergonomics (i.e., concerning mental processes), and organizational ergonomics (i.e., optimization of socio-technical systems, including their structures and processes). HFE is a multidisciplinary, human-centered integrating science that is not domain-specific (ILO and IEA, 2021). HFE encompasses not only safety and health but also the cognitive and psycho-social aspects at work. Additionally, HFE can focus on micro ergonomic design aspects—including the design of the procedures, the context, and the equipment and tools used to perform tasks—as well as macro ergonomic design aspects—including the work organization, types of jobs, technology, work roles, and communication (Wilson, 2014). Through their work activity, human beings acquire experiences about physical and social, external as well as internal reality, and they change by self-reflection (Schön, 1983).

A socio-technical "work system" is part of a work process in which a task is accomplished. Through the interaction of working humans with work equipment, the function of the system is fulfilled within the workspace and work environment under specific working conditions (ISO 6385, 2016). "Work design" is a collective term for measures for the purposeful and systematic design of work objects, work processes, and working conditions. The aim of work design is the optimal fulfillment of work tasks, taking into account human development requirements and economic efficiency (Dyckhoff, 2006). In order to increase productivity, work design is based on the rational principle of labor division and the systematization of work processes. The division of labor supports specialization, which makes it possible to use machines and automate work processes. Work design determines which functions are automated and which remain with humans (Baxter and Sommerville, 2011). The systematization of work sequences aims at their method-based optimization (Schlick et al., 2018). Since the division of labor goes along with external supply and a loss of autonomy, the social dimension of human-centered work seems to be indispensable; it is expressed, among other things, in committed cooperation, fair work relationships, and learning opportunities (Ulich, 2011).

2.2. Human-centered work design and interactive human-technology systems

Human-centered work design refers to a problem-solving approach from a human perspective and the interactions of the working human with technical and organizational factors. One application focus concerns interactive human-technology systems (i.e., interaction design). For this purpose, knowledge and methods from human factors, ergonomics, usability, and user experience are applied. Human-centered design criteria are documented in legal regulations, standards, and rules (Karwowski et al., 2021).

Human-centered work is characterized by balanced workload situations in order to avoid over-and under-strain and thus to promote performance, job satisfaction, and learning (Schlick et al., 2018). Work activities should not harm the health of the working person, should not-or at most temporarily-impair their wellbeing, meet their needs, enable individual and collective influence on working conditions and work systems, and contribute to the development of their personality in the sense of promoting their capability potentials and competencies (Ulich, 2011). Design dimensions include work content, time, process, conditions, workplace, or equipment.

Concerning the design of interactive human-technology systems, ISO 9241-2 (1992) Part 2 specifies the human-centered requirements of user orientation, variety, holism, meaningfulness, the scope for action, social support, and development opportunity. In ISO 9241-110 (2020) Part 110 (interaction principles), the human-centered design criteria of task appropriateness, self-describability, expectation conformity, learnability, controllability, error robustness, and user binding of technical systems are concretized. ISO 9241-210 (2019) Part 210 (Human-centered design for interactive systems) emphasizes the user experience, which describes the functional and emotional impressions of a user when interacting with a product or service. If users' requirements are met in a useful way, it is pleasant to use a product.

2.3. Human work and human factors

Technical rationalization measures aim to raise work productivity and reduce costs also by substituting the human workforce with machines. Such automation efforts can affect both physical and mental work. To automate sophisticated forms of work that were previously preserved by humans, the use of AI aims to imitate human intelligence (Landgrebe and Smith, 2022). In order to assess the extent to which AI can explain, predict, and influence human behavior requires HFE expertise.

Humans represent a physical, biological, social, and mental entity. They are generally aware that they have a perception, a mind, and a will. A basic requirement of human existence is to access the natural and social environment through interpretive understanding to gain meaning. In communicating with other humans and interacting with the environment, humans gain a deep understanding of the world (Bender and Kolle, 2020). On this basis, they can act purposefully even in the face of incomplete and contradictory information (Wilkens et al., 2014), transfer existing experiential knowledge to new contexts through the understanding of meaning, adopt different perspectives, and anticipate the consequences of their actions. However, the world can only be understood through knowledge of concepts and symbols that emanate from human intelligence (Whorf, 1963). Human intelligence is a complex neurological capacity including reasoning, memory, consciousness, emotions, will, intention, and moral judgment. Humans can take responsibility for acting solidaristic and morally (Böhle, 2009). Such subjective and implicit competencies are the basis of knowledge and innovation work:

- "Knowledge work" refers to the execution of work orders that are to be carried out according to available (un-) complete rules that the working person knows. Knowledge includes

explicit or implicit components that guide action but cannot be verbalized.
- "Innovation work" refers to problem-finding and problem-solving work assignments in which the goal and the path to achieving the goal are not predetermined. Innovation work includes unplannable or poorly plannable, unpredictable intellectual performances, prognostic intellectual performances that do not have precisely defined goals, and diagnostic intellectual performances for which no algorithms can exist because it is unclear what is being searched for in the first place (Hacker, 2018).

There is increased interest to substitute knowledge work by AI. However, as expected, significant aspects of innovation work will remain with humans in the future (Hacker, 2018).

2.4. HFE design criteria for HCAI

HCAI encompasses individual safety, the trustworthiness of AI operation, an appropriate division of functions between humans and machines, and conducive working conditions (Huchler et al., 2020). Insofar as AI applications contribute to the automation of knowledge work, the design criteria of conduciveness to learning and social compatibility take on increased importance.

HCAI ties in with the human capabilities and places the mutual reinforcement of humans and AI at the center of their interaction. Such reinforcement will be reached through a complementary division of functions between humans and AI that takes into account the differences between human capabilities and technical functionalities (Rammert, 2009). The core principle of the complementary division of functions strives for higher productivity and adaptivity through a lower overall degree of automation with increasing partial automation and systematic integration of the capabilities of the working human (Huchler, 2022). In order to cope with the uncertainties of work systems, human options of control are extended, e.g., by informal work actions and work-integrated learning processes. Appropriate competencies are to be maintained by designing the interaction between humans and AI in a learner-friendly way (Grote et al., 2000).

Regarding the social responsibility of AI applications, ethical aspects need to be clarified. Ethics focuses on specifically moral action, especially with regard to its justifiability and reflection (Bostrom and Yudkowsky, 2013). In AI use, ethical questions are concretized in the phenomena of uncertainty and social inequality (Brynjolfsson and McAfee, 2012):

- *Uncertainty:* Purposes of the use of technology are not or not exclusively achieved, i.e., the relationship between means and ends is not always comprehensible; this manifests in insufficient transparency, comprehensibility, and manageability, as well as the irreversibility of decision-making processes.
- *Social inequality:* The people who suffer harm are not the beneficiaries; inequality affects personal and informational autonomy in the use of data, the possibility of personal development, decision-making power, and the economic exploitation of work results.

When designing HCAI systems, ethical rules should ensure that human autonomy of decision and action is preserved with moral intent (Floridi et al., 2018). Currently, AI is not yet comprehensively capable of making moral decisions. Moral principles are instead specified by humans and implemented in the form of algorithms that can lead to morally grounded actions as a result. An essential principle of human-centered design is preserving and appropriately using these moral means of control and access (Bülchmann, 2020). This also relates to the human influence of exit points, if necessary.

Bostrom and Yudkowsky (2013) recommend four ethical design principles of HCAI: An HCAI functioning should be (1) comprehensible and (2) its actions predictable in principle; there should be sufficient time for users to react and veto control in the event of a potential malfunction. HCAI should (3) not be easily manipulated, and if a malfunction does occur, (4) responsibility should be clearly defined.

HFE traditionally incorporates a variety of perspectives on human work. In this chapter, many methods and definitions of HFE were presented, which are also relevant to the other disciplines. HFE has a pragmatic concept for the design of work systems, processes, and tools, including AI, some of which are documented in regulations and standards. Knowledge of these methods and definitions is cross-disciplinary relevant, and not limited to HFE, and additionally necessary to prepare the implementation of HCAI at work.

3. Information is the key: HCAI in information science

The topic "human-centered AI" has not gained much attention in information science; even "human-centricity" is not a much-discussed concept due to how information science's central object of study — information — is conceptualized and defined. In the following, it will be argued that there is no need to be explicit about "human-centricity" because, for one, information does not exist outside of human beings, and second, information behavior is central to the development and evolution of humans. If human-centered AI systems are considered information systems, then several implications can be deduced for their design and handling from this argumentation. Overall, there are three ways how information science can conceptually approach human-centered AI:

(1) The meta-level that discusses HCAI against the discipline's pragmatic understanding of information.
(2) The information behavior perspective, which is considered central to human life and that leads to the creation of user models.
(3) The literacy aspect that reflects on the skills humans need to handle AI successfully.

3.1. The meta-level: what is information?

In the literature, *information* is often characterized by using the Semiotic Triangle by Charles Sanders Peirce since it has been argued that information is the basis for the communicative action (in the sense of information as a message, information

as a state) and a communicative act itself (in the sense of exchanging knowledge/being informed, as a process; Henrichs, 2014). Accordingly, information has a triadic structure consisting of the following:

- *Object*, meaning of the information or semantics,
- *Sign(s)/syntax*, signals for or physical carrier of knowledge and formal-syntactic representation of signs, and
- *Interpretant*, user/usage of information, or pragmatics.

This structure is dynamical since forming the relations between the parts always entails some sort of transmission. The structure is also relational, which results in the need to consider all three parts of the triangle simultaneously when referring to *information* since they are linked inseparably. Information science deals with all three aspects of information, leading to multi- and interdisciplinary studies with, e.g., computer science (that mostly focuses on the signal part, for example, when building digital libraries) or linguistics (focusing on the object part, for example when constructing ontologies or other knowledge organization systems). The most fruitful — and therefore most central — avenue for information science concerns, however, the pragmatics of information, which focuses on the human part of information processing and how humans make use of information. In a nutshell, information science seeks to understand how and for what reasons humans need, gather, and use information. It primarily asks from the interpretant point of view: What is information used for? Which actions are possible with that information, and what do humans need to act properly?

This understanding of information differs from the definition used in, e.g., computer science or telematics that focuses on the signal and disregards the meaning of the information (Shannon, 2001). Information science considers knowledge as the raw material for the creation of information — knowledge is possible information (Rauch, 1988), and information is a manifestation or representation of knowledge (Kuhlen, 2004). Since (formless) knowledge, which exists independently from signals, needs to be brought into a (physical) form to be transmitted (Stock and Stock, 2015), it can be argued that "information is a thing — knowledge is not" (Jones, 2010). If humans use that potential information for further action, information materializes. In general, information is used to decrease the amount of uncertainty a human experiences (Wersig, 1974).

Therefore, information depends on the context in which humans perceive it, and it can be different with different contexts and different humans. Humans construct information by decoding the signal — information does not just exist (whereas knowledge exists even without signals; Stock and Stock, 2015). This construction takes place in social environments and via means of communication. The recent popularity of ChatGPT and generative language models for AI, as well as how interaction (or communication) with those systems has been designed, is reflective of the relevance of communication in information processes and for information behavior.

3.2. Information behavior and user modeling: traces toward the human perspective

Reflections on the concept of *information* and studies on how humans engage with information have mutually influenced each other, developing a shared understanding of the subject. Information behavior research, as a sub-discipline of information science, is concerned with how and why humans interact with information in different informational contexts, including how they use, create, and seek information (Bates, 2017), actively or passively (Wilson, 2000), individually, collectively, or collaboratively (Reddy and Jansen, 2008). Information non-use, such as information avoidance (Golmann et al., 2017), is also part of the research agenda as well as information sharing, (personal) information management, information practices, information experiences, and information discovery (Greifeneder and Schlebbe, 2023).

It is remarkable that, similarly to the triadic structure that considers the human, the interpretant, an inseparable part of *information*, the concept of *Human Information Behavior* was never adopted by the research community (Bates, 2017). Information behavior research underwent several so-called conceptual and methodological turns that are also reflective of the increasing relevance and attention the human being has been attributed over the years: from understanding which information sources and systems humans use to gather information, learning about the information need that motivates humans to interact with information (cognitive turn) and their emotions involved (affective turn), to the role of socio-cultural contexts (socio-cognitive turn) and habitualized information practices (social-constructionist turn) (Hartel, 2019). Information behavior can only be exposed by humans — manifesting their (for outsiders' implicit) relevance for information science again.

This also becomes apparent by one central activity of information behavior, i.e., the humans' engagement in looking for information. Case and Given (2016) emphasize that from birth onwards, humans are prompted to seek information to meet their fundamental needs. Information needs are driven by those fundamental needs, often because the human recognizes a lack of information to meet the fundamental needs (Stock and Stock, 2015). Information seeking behavior is activated by concrete information needs and is, therefore, active and intentional (Case and Given, 2016). If a computer or IT system is used to look for information, then Wilson (1999) speaks of Information Searching.

Information behavior research is not a goal in itself — like other disciplines, it seeks to advance information systems to tools that can be easily and efficiently used, that automatically adapt to changing situations, and that are adaptable to the needs of their users (Elbeshausen, 2023; Lewandowski and Womser-Hacker, 2023). As has been argued, understanding information and information behavior always requires knowledge about humans, e.g., users of an AI system. The complexity of information behavior often prevents the use of quantitative or statistical methods, so that qualitative methods are the main approach.

User modeling is an important activity in this regard, which aims to describe individual users to enable, for example, personalization of search results or groups of typical users that share certain characteristics (e.g., novices and experts). Latter is often realized via personas that represent typical users of a system with very concrete properties (an approach sometimes criticized for replicating stereotypes; Marsden and Haag, 2016). If the information system targets a broad user base, user modeling can be a tough challenge since there is not only a large, diverse group of (possible) users, but user behavior is also dynamic and can change while using the system for particular tasks or over time. Humans adjust informational practices, tactics, and behavior dynamically to match contexts and to maximize the amount of information they can get, e.g., by changing search terms (Pirolli and Card, 1999). In addition to the informational environment and contexts, the information behavior of a person is also connected to their personality (Lewandowski and Womser-Hacker, 2023). The principle of least effort (Zipf, 1949) is also applicable to information behavior: humans tend only to spend the minimum effort to accomplish tasks, often only resulting in only a satisfying (but not the best) result.

Interactive information systems that more dynamically react to users' information behavior and that serve a variety of tasks, need to even better understand humans, their needs, and context to be accepted and add value. Ingwersen and Järvelin (2005) argue — similar to the pragmatic definition of information — that information systems are never used in isolation but are always embedded in personal, organizational, cultural, and more contexts and therefore need to be designed and evaluated within those contexts. Users should be given the opportunity to use information systems purposefully to focus on the task to be fulfilled without being bothered by the challenges of handling the system (Elbeshausen, 2023, p. 474). For economic reasons, in the corporate context, it is of paramount importance to know which information types and information services are meaningful for employees and which information needs arise (Stock and Stock, 2015). Gust von Loh (2008) distinguishes between objective information needs from workers and employees that arise from a certain job position (e.g., because of the company strategy) and that are independent of a particular staff member, and subjective information needs that are articulated by a specific job holder and that stem from user studies.

Here, strong connections to the human-computer interaction (HCI) field become apparent. Information science and HCI share their focus on humans interacting with information systems, their cognitive and contextual embeddedness while doing so, and the subjective and objective information needs a system has to satisfy (Jetter, 2023). Both disciplines acknowledge that the design of humane (or human-centered) information systems (empirically proven via usability-and user experience methods) benefits from the enrichment of information and contextualization.

Technical information systems may be unable to fit all the information behavioral aspects of a broad user base but may need to focus on a selection of tasks or user types. Furthermore, information behavior also takes place outside of technical or digital environments. Then, educating the users toward a certain behavior and increasing their knowledge about the information system and environment could be an additional approach.

3.3. Information literacy as a prerequisite to deal with AI

Although information is central to human development and life, dealing with information in a good and meaningful way is a skill that has to be acquired and cultivated — especially with regard to the ever-increasing complexity of today's digital information environments. To be able to efficiently and ethically deal with information in a particular context, to understand how information is produced, evaluated, and distributed, how it can be effectively searched for, and to assess the personal informational and thinking competencies critically are skills that are subsumed under the term "information literacy" (Griesbaum, 2023). The UNESCO (2013) considers media and information literacy as a core competency for democratic societies that enables citizens to successfully engage and participate in private, vocational, and societal activities. Information literacy is, however — and similarly as the concept at its core: information — a relational concept. Its characteristics change with the information environments and contexts in which human beings have to deal with information (Griesbaum, 2023). This also presumes that an information-literate person has a certain amount of knowledge about the topic or circumstances they are dealing with — but this is not always the case. Hence, the more the person lacks expertise and knowledge, the more trust the person needs to put into the information ecosystem. Information literacy then transfers from the topic or situation itself to the evaluation of other information sources or experts whose recommendations have to be trusted (Griesbaum, 2023).

Despite its stated relevance, often, information literacy is not an integral part of school education but rather embedded in higher education and services of university libraries (ACRL, 2016). In work-related contexts, information literacy issues become apparent in enterprises with structured knowledge management approaches (Travis, 2017). However, Lloyd (2013) has found that information literacy at the workplace is mainly reduced to socio-cultural practices for collaboration. Middleton et al. (2018) could show that information literacy is strongly connected to innovative work practices.

This hints toward an increasing need for information literacy in complex information contexts as induced by AI systems. Consequently, AI literacy is an emerging field in information science, borrowing most of the central aspects already embedded in information literacy but also highlighting further skills and normative claims (Touretzky et al., 2019). Ng et al. (2021a,b) performed a literature search on AI literacy to derive aspects this concept entails. They found that all selected articles consider knowing the basic functions of AI and how to use AI applications in everyday life ethically (know and understand AI), as well as applying AI knowledge, concepts, and applications in different scenarios (apply AI), the core competencies of AI-literate humans. Two-thirds of the analyzed articles also mention critical higher-order thinking skills (such as evaluating, appraising, predicting, and designing) as part of AI literacy (evaluate and create AI).

Furthermore, the literature states that AI literacy is central to the future workforce, simultaneously preparing humans to efficiently use and critically evaluate AI and sparking career interest in this field (Chai et al., 2020). Interestingly, the study revealed

that only 50% of the articles considered educating humans about socially responsible behavior when using or designing AI as part of AI literacy. Here the authors see room for improvement: "[…] conceptualizing AI literacy with human-centered considerations is crucial to building a future inclusive society" (Ng et al., 2021a, p. 507). The evaluation of AI literacy itself is conducted via knowledge tests, self-reporting, questionnaires, or observations when interacting with AI.

4. Designing learning opportunities for all: HCAI in adult education research and adult learning

The core assumption of HFE is that human beings are able and willing to learn and shape their working lives. Considering this first assumption, aspects of lifelong learning, adult and continuing education, and adult learning at the workplace touch the core of work design and human centeredness. At the same time, for many years adult education policies and research have dealt with how educational systems can effectively provide knowledge and skills for a technologically changing (working) society (Merriam and Bierema, 2013). The results of numerous research activities within the adult education scientific community contribute to shaping AI-affected workplaces in a human-centered way, which from an adult educational perspective means a learning-centered way (Harteis, 2022).

From the perspective of adult education research and policy, it is a consensus that educational systems target fostering the quality of educational processes and providing learning opportunities equally (UNESCO, 2019; BMAS and BMBF, 2021; Council of the European Union, 2021; Autor:innengruppe Bildungsberichterstattung, 2022; OECD, 2023).

Educational systems contribute to designing and establishing *learning opportunities*. That means they contribute to channeling, organizing, and monitoring informal, non-formal, and formal (adult) learning processes. In democratic states, educational systems aim to provide skills to individuals so that they can actively participate in public and working life. At the same time, quality learning processes within educational systems underlie the expectancy to provide a qualified and employable workforce. Scientific discourse treats these aspects using the two concepts of individual self-regulation, on the one hand, and human resources, on the other hand (Autor:innengruppe Bildungsberichterstattung, 2022).

Work has a major role to play in education. First, work and working life are significant fields of adult learning. Human subjects acquire skills for and within their employment to stay employable. Second, in Western-so-called working societies-work is a major part of active participation in society, as it impacts, e.g., social status and social and financial resources as much as professional and, therefore, social identity (Kraus, 2008; Gericke, 2017). Third, work in terms of work-based learning is a learning and teaching methodology (Bauer et al., 2004; Dehnbostel, 2022). Therefore, it is an important quality criterion of professionally designed adult and continuing education to take aspects of individuals' (working) lives, such as possible ruptures in (working) biographies and career development, as one starting point for developing and creating learning opportunities.

Equality of opportunities in adult education refers to equal access to learning opportunities as a major challenge for education systems (Käpplinger and Lichte, 2020; Council of the European Union, 2021). It targets especially vulnerable and marginalized groups such as migrants, low-qualified, unemployed, disabled, or illiterate persons, who have different learning needs regarding content but also need differently structured learning opportunities than high-skilled workers or middle-class citizens. Work has an important role to play in the equality of learning opportunities. According to the Adult Education Survey (BMBF, 2022), for years, more than 70 up to 75% of the adult learning activities of the German population aged 16–65 have taken place during daily working time or were financed by the employer. A much smaller and even decreasing part of adult learning activities was work-related but based on individually generated financial and time resources (13% in 2012 – 8% in 2020), while the share of individual non-work related learning activities is relatively stable at about 17–18% (BMBF, 2022, p. 22). Major differences exist in the participation rates of different social groups. The employed population shows a higher participation rate (46% in 2012 – 60% in 2020) than the unemployed population (13% in 2012 – 19% in 2020). The same applies when comparing the un-or low-skilled population (30% in 2012 – 46% in 2020) with high-qualified persons (about 70% in 2012 – 81% in 2020). Remarkably, in the German adult population, learning activities are on the rise in absolute numbers. At the same time, the differences in share between certain social groups have not remarkably diminished. These findings concerning the participation rates in adult learning vary across countries, still the gap between employed and unemployed, as much the high – and the low-skilled persons, shows to be a central challenge in more or less all OECD countries (European Union, 2021).

Against this background, an adult education research perspective on human-centered AI implementation will concentrate not only on how to provide quality learning opportunities but it will focus as well on how to tailor these quality opportunities for each social group. So, an adult education research perspective contributes to human-centered AI, first of all, by analyzing if an educational system, an employer, or a single workplace offers learning opportunities for AI-based workplaces, whom these learning opportunities are made for, and what kind of learning opportunities are proposed.

4.1. Designing learning: what skills should we qualify for?

Regarding contents and needs for skills, there is consensus that in a digitalized and AI-based world, life and work tasks will become more complex. There are catalogs trying to capture and describe important future skills. In the context of education and lifelong learning, the European Commission's Framework DigComp has had quite an impact in the field in Germany (Joint Research Center (JRC), 2022). Moreover, in higher education, the so-called twenty first-century skills play an important role (Anandiadou and Claro,

2009; Schnabel, 2017). These two catalogs represent important examples for a whole discussion that brings the importance of future skills to the fore. They concentrate on skills in

- Working with media, technology, information, and data
- Virtual and face-to-face communication and collaboration in diverse (e.g., interdisciplinary, intercultural, intergenerational) contexts
- Creative problem solving, innovation, analytical and critical thought
- Flexibility, coping with ambiguity, self-motivation, and working independently (Schnabel, 2017)

When thinking human centeredness from an adult learning perspective, it is important to note that these skills will not replace professional skills but will additionally be on top of professional skills. Even more, they will be interlinked with professional skills. So when preparing a workforce for an AI working world, degrees will have to encompass professional skills as a basis plus these future skills.

4.2. Designing learning: who should we target and how?

Quality is not determined by knowing the skill needs and contents of learning but also by methodologies that help to teach these skills professionally, effectively, and efficiently to a whole range of target groups. When it comes to skill delivery in companies, there is a vivid research landscape on how to deliver sustainable learning success in digital transformation. The learning and teaching methodologies in focus range from informal learning in the workplace, learning nuggets, non-formal workshop settings, or formal learning arrangements within chambers and universities (e.g., Rohs and Ganz, 2015; Anderson and Rivera-Vargas, 2020). From the company's viewpoint, where there are financial and economic restrictions, these discussions are critical. With a company's decision to invest in one or another kind of learning opportunity, it shapes structures and methodologies of learning and, finally, participation rates in adult learning opportunities to a high degree.

Discussions of teaching and learning methodology differentiate along the question of which knowledge or skill can be efficiently and professionally taught in which setting to which target group. Taking marginalized groups as an example, it is a common educational argument based on Bourdieu (Watkins and Tisdell, 2006) or biographical research (Alheit, 2021) that low-skilled people or functional illiterates have rather negatively experienced learning throughout their lives, sometimes they have gone through biographies of failing in an educational system. Therefore, it is highly challenging for professionally organized adult education to get access to these groups and to teach them effectively-much more challenging than teaching high-skilled people or managers who have had successful learning careers.

Therefore, when implementing AI in a human-centered way, quality learning opportunities need to ensure that all target groups who are affected by AI in the workplace get the opportunity to qualify for these changes. At the same time, different target groups will need different skills in the workplace and different learning methodologies for acquiring these skills. In addition, it is a professional adult education task to create good learning opportunities with a well-fitted methodology that facilitates between the affordances of a company within the digital transformation, on the one hand, and the needs of the target group and their learning habits, on the other hand.

In terms of learning methodologies, recent projects have shown

(a) That professionally implemented learning projects in the workplace can effectively qualify low-skilled workers on the job within digital transformation processes (Goppold and Frenz, 2020).
(b) That worker's councils have an important role to play as facilitators of bringing together unskilled and low-skilled workers or functionally illiterate employees with continuing education activities. Still, the members of worker's councils need to be qualified to fulfill their role (Lammers et al., 2022; Arbeiter, 2023).

In recent years, networked structures in adult education have been brought to the fore (e.g., UNESCO, 2015). In the case of implementing human-centered AI networks between adult education providers, companies and worker's councils will probably be in favor of channeling professionally tailored learning opportunities into companies. Taking Germany as an example, vocational training providers create those networks in order to target marginalized groups; in the case of high-skilled individuals, universities of applied sciences have a mandate of developing continuing education to create opportunities in cooperation with companies (Dollhausen and Lattke, 2020). It is an issue if these networks allow scaling up learning opportunities for a whole population or multiple companies and not to tailor adult learning for one single company.

5. Technology, autonomy, and trust: HCAI in psychology and human-computer interaction

Typically, the AI research community focuses on algorithmic advances, deeming a human-centered approach unnecessary, but at the same time, human-centered thinking is gaining popularity, and the AI community is diverse. However, this new thinking challenges established practices (Shneiderman, 2022, p. 40). This new thinking also influences the perception of psychology and HCI, which are closely related, although they are separate disciplines, because psychology plays a significant role in HCI. For this reason, both disciplines are discussed together in this section (Clemmensen, 2006). From a psychological and human-computer interaction (HCI) point of view, technology acceptance and adoption are also becoming essential aspects of human-centered AI (Del Giudice et al., 2023), especially in human autonomy (Bennett et al., 2023) and the development of guidelines for human-AI interaction (Amershi et al., 2019). Such guidelines need to consider issues with information overflow and should assist in using complex

systems (Höök, 2000). This consideration can be achieved by putting in place verification measures or regulating levels of human-controlled autonomy to prevent unintended adaptations or activities by intelligent systems (Amershi et al., 2019; Xu et al., 2023). In addition, AI-driven influence techniques like psychological targeting or digital nudging have raised ethical worries about undermining autonomy (Bermúdez et al., 2023). Moreover, a series of recent studies found that employees who work with AI systems are more likely to suffer loneliness, which can lead to sleeplessness and increased drinking after work (Tang et al., 2023).

Nevertheless, in HCI, the understanding of human autonomy remains ambiguous (Bennett et al., 2023). This ambiguity might be attributed to an old controversy if people and computers being in the same category or if, as many HCAI sympathizers believe, vast differences exist (Shneiderman, 2022, p. 25), with Shneiderman supporting the latter (Shneiderman, 2022, p. 31). However, AI and its impact on the workplace are said to be disruptive, including chatbot-based communication systems that can demonstrate empathy through an understanding of human behavior and psychology, allowing the chatbot to connect with customers emotionally to ensure their satisfaction and thus support the adoption of AI systems (Krishnan et al., 2022). AI differs from HCAI by two key human-centered aspects in terms of performance and the product. The human-centered process is based on user experience design methods and continuous human performance evaluation. Furthermore, the human-centered product is emphasized by human control to enhance human performance by designing super tools with a high level of automation (Shneiderman, 2022, p. 9).

Still, HCI acknowledges the importance of AI by highlighting it in almost all of the current HCI grand challenges, like human-technology symbiosis and human-environment interactions, to name a few (Stephanidis et al., 2019). This is accompanied by six grand challenges of human-centered AI: human wellbeing, responsible design of AI, privacy aspects, AI-related design and evaluation frameworks, the role of government and independent oversight, and finally, HCAI interaction in general (Garibay et al., 2023). HCAI interaction especially plays a vital role at work, as economic challenges meet with ethical and organizational considerations (Garibay et al., 2023). This collection of grand challenges reflects the almost symbiotic relationship between HCI and AI.

Finally, the transition to human interaction with AI systems by moving on from siloed machine intelligence to human-controlled hybrid intelligence can be considered a new opportunity for HCI professionals to enable HCAI (Xu et al., 2023). A potential goal of human-centered AI design is to create human-controlled AI using human-machine hybrid intelligence, which emphasizes the integration of humans and machines as a system, aided by the introduction of human functions and roles that ensure human control of the system (Xu et al., 2023, p. 503). However, such integration of humans and machines is not without obstacles and unrealistic user expectations, and negative emotional responses are often a source of concern.

5.1. Unrealistic user expectations

Exaggerated and unrealistic user expectations about AI-based applications and absent design solutions to support human-centered work can lead to frustration and questioning the "intelligence" of such applications (Luger and Sellen, 2016). For example, high efficiency of search functions should be combined with curated content and meaningful recommendations even without the necessary meta-information. These demands raise hopes that may neglect the actual software and hardware capabilities of research projects, which may only be feasible for very large companies. In addition, it is requested to combine, match, and recommend different kinds of heterogeneous data, even on the internet, without considering resources. Although AI makes significant improvements daily, these are still quite unrealistic user expectations today.

One possible way to overcome these challenges of unrealistic user expectations is to divide the AI-based processes into different phases. For example, Amershi et al. (2019) offer 18 AI design guidelines separated into four phases: the initial phase when beginning to work with an AI-based application, during general interaction, when things go wrong, and aspects considering long-term experiences. A vital aspect of these guidelines is providing support and managing expectations. Such aspects are not unknown to technology acceptance models.

5.2. Building on technology acceptance and trust

Technology acceptance models, e.g., TAM (Davis, 1989), UTAUT (Venkatesh et al., 2003), and their extensions in various fields (Kao and Huang, 2023), offer a promising domain for an evaluation concerning human-centered AI. TAM is a conceptual model used to account for technology usage behavior, which has been confirmed to be valid in various technologies among different groups of people (Venkatesh et al., 2003; Choung et al., 2023). The original TAM model postulates that the intention to use technology in the future is determined by two key factors: perceived usefulness and perceived ease of use (Davis, 1989).

Choung et al. (2023) integrated trust as an additional variable in their extended TAM model. Their two studies confirm that trust is vital for accepting technology. Therefore, AI technologies should be designed and implemented in a human-centered way; consequently, their implementation should be easy to use, useful, and trusted. In general, empirical findings support the assumption that technology acceptance models help to explain the acceptance of AI technologies (Sohn and Kwon, 2020), including the aspect of trust (Choung et al., 2023). However, nevertheless, there are limitations to their usage, which are discussed by Bagozzi (2007).

Users' low level of trust in how their data is handled and processed must also be adequately considered psychologically. AI, in general, can predict user behavior in a wide range of applications by following digital traces of usage. Besides legal and ethical challenges, psychologists call this approach digital phenotyping when using elaborated smart sensing techniques and when it is successfully assisted and analyzed by data mining and machine

learning tools (Baumeister et al., 2023). This is not an entirely new topic, as, e.g., user behavior in an online environment relates to their personality and can be used to tailor content, improve search results, and increase the effectiveness of online advertising (Kosinski et al., 2014), which is backed by many empirical studies and summarized by Baumeister et al. (2023). At the same time, ethical challenges are addressed by the human-in-the-loop design, where individuals are asked to make a final decision or action (Shneiderman, 2022; Garibay et al., 2023; Xu et al., 2023), which can also help to improve the trust to HCAI.

This is in line with results from an experiment by Westphal et al. (2023), in which they empower users to adjust the recommendations of human-AI collaboration systems and offer explanations for the reasoning of the systems. The idea behind this approach is to counter low trust and limited understanding of users dealing with recommendations of an AI system, and at the same time, to keep in mind to achieve an adequate or calibrated trust, meaning that, e.g., not to over trust the AI system (Leichtmann et al., 2023b). However, interestingly, explanations could backfire because they can increase or signal task complexity, whereas enhanced decision control leads to higher user compliance with system recommendations (Westphal et al., 2023). These results affirm that well-explained support can be essential to accept and facilitate HCAI at work, leading to HCAI systems that explain themselves, so-called human-centered explainable AI, which can be accompanied by educational offers and measures for providing human-centered explainable AI.

5.3. A glimpse into the near future: explainable, understandable, and gamified AI

Such explainable AI, especially if it is human-centered, can be considered crucial. However, from a socio-technical standpoint, AI should also be understandable to stakeholders beyond explainability (Habayeb, 2022). This can be achieved when implementing user-participated experimental evaluation because it is necessary to overcome the relatively simple unilateral evaluation methods that only evaluate AI systems' performance.

One way to implement this is to use, e.g., a gamified crowdsourcing framework for explainability (Tocchetti et al., 2022), which uses game design elements in a non-game context (Deterding et al., 2011). However, current research focuses primarily on strategic and system issues related to AI system performance (Raftopoulos and Hamari, 2023), which limits the view of AI. Furthermore, HCI should promote the evaluation of AI systems as human-machine systems by including the end-user perspective (Xu et al., 2023, p. 505). Nevertheless, this makes it necessary for HCI to enhance its current methods. Constraints like focusing on single user-computing artifacts with a limited context of use, lab-based studies, or static human-machine functions are prevalent. Instead, "in-the-wild studies," the application of distributed contexts of use, and longitudinal study designs are encouraged to address the identified unique issues of AI systems to influence the development of AI systems in a human-centered way (Xu et al., 2023, p. 509–512).

Additionally, incentives for using artificial intelligence, e.g., gamification, are also emerging topics of interest for human-centered AI (Mazarakis, 2021). Gamification tries to bring the motivating effect associated with games to non-game situations with the help of elements like badges, leaderboards, and points (Mazarakis, 2021, p. 279, 283). First studies conducted with intelligent user interfaces and voice user interfaces like Amazon Alexa, which are also considered social robots and active appliances in artificial intelligence (Shneiderman, 2022), show the potential to focus on empirical research for the acceptance of these interfaces (Bräuer and Mazarakis, 2022a,b; Haghighat et al., 2023). Nevertheless, explainable AI is also, in this use-case, key to counteracting suspicion regarding the trust of social robots and active appliances in artificial intelligence. For example, to achieve transparent and accountable conversational AI and to include such a system in a gamified environment, interpretability, inherent capability to explain, independent data, interactive learning, and inquisitiveness are necessary (Wahde and Virgolin, 2023, p. 1856). Inquisitiveness is meant to be by the AI to show curiosity and not to annoy the user to achieve human centeredness. Curiosity means just displaying inquisitiveness in specific contexts, such as during learning, so as not to disturb the user (Wahde and Virgolin, 2023, p. 1865).

A further step is taken by Tocchetti et al. (2022), which propose a gamified crowdsourcing framework for explainability. Their crowdsourcing framework engages users on different levels than other platforms, primarily relying on extrinsic rewards. The provided user education, in particular, would raise users' understanding of the types of information that an AI system requires, learns, and produces, improving users' efficiency and developing the users' mindsets (Tocchetti et al., 2022, p. 7). Furthermore, their work shows that a symbiosis of HCAI, gamification, and explainable AI is also possible with greater effort and exertion. Consequently, this also results in an increased human centeredness. A first effort of studies in game-based environments yields promising results for explainable AI (Leichtmann et al., 2023a).

Different scenarios for using gamified AI and gamification in the context of AI, in general, are possible. For example, Tan and Cheah (2021) describe a work in progress and prototype for developing an AI-enabled online learning application for lecturing at a university physics. However, this scenario can be switched to a work-related setting without much effort. As education is one of the main application areas of gamification (Mazarakis, 2021), a combination with AI is obvious and already taking place (Kurni et al., 2023). In this case, first data is collected for AI processes, e.g., through step-by-step scaffolding instructions and feedback to students by studying students' progress in answering quiz questions. Then, it is possible to implement adaptive assessments to more accurately identify the student's level of mastery, adjust the difficulty level and the number of questions at each level of difficulty, and finally, step between each level of progression based on the student's answer. Thereby, individual feedback, which can then be used for learning analytics to improve, optimize, or redesign the curriculum to meet the needs of specific student cohorts, can be provided (Mazarakis, 2013; Tan and Cheah, 2021), e.g., for employer-provided training in different work scenarios.

6. Discussion

This chapter presents conclusions from the previous chapters' theory and practice and shows relations between them in order to inform a synergistic human-centered AI theory. First interdisciplinary human-centered AI perspectives are shown, according to Wilkens et al. (2021), and how they relate to the five disciplines. Then, interdisciplinary views of human-centricity and their interrelations are matched with observations from a demonstrator. These views have the goal of setting foundations for an interdisciplinary synergistic theory of human-centered AI, which is presented in Section 6.3.

6.1. Interdisciplinary perspectives on human-centered AI

Wilkens et al. (2021) found five co-existing views in a comprehensive literature review analyzing the significance of HCAI: a deficit-oriented, a data reliability-oriented, a protection-oriented, a potential-oriented, and a political-oriented understanding of how to achieve human-centricity while deploying AI in the workplace. These five perspectives reflect many aspects of AI's human-centricity, with varying levels of maturity along each dimension. In order to put the results of this article into context, the disciplines of information science, human-computer interaction, psychology, and adult learning are related in Table 1 to the five perspectives of Wilkens et al. (2021). HFE is inherent in Wilkens et al. (2021) and would cover all five perspectives, so it is not shown in Table 1.

The perspectives of Wilkens et al. (2021) largely coincide with the results of the demonstrator. Although all the perspectives are covered, HFE is particularly concerned about the data *reliability-oriented understanding* of HCAI and the *potential-oriented understanding*.

Potential deficits of *data reliability* are mainly considered from an ethical perspective since technical artifacts are not ascribed to any moral competence; at best, they can imitate human moral behavior. Taking into account the instrumental character of AI, it is rather necessary to consider fundamental conflicts of interest that might favor immoral behavior. This, however, leaves the field of AI design.

The *potential-oriented understanding* promotes a hybrid design approach corresponding to the "complementary division of functions between humans and AI." The mutual reinforcement of humans and HCAI appears to be suitable for satisfactorily coping with future, currently potentially unknown working requirements. It emphasizes the evolutionary principle of humans, whose behavior is imitated by intelligent machines, and that will ultimately be reflected in the precision and reliability of machine procedures.

The information science perspective on human-centered AI can be summarized mainly as *protection-oriented*, *potential-oriented*, and *political-oriented*. It is *protection-oriented* when it studies how information systems should be designed so that humans can easily and safely use them, and it is *potential-oriented* since it considers information systems sociotechnical environments in which humans co-construct information with the help of technology. In social settings, and especially at the workplace, the *political-oriented* perspective is also part of information science's agenda, e.g., in terms of information literacy.

An adult education perspective, which in its tradition always includes an advocacy perspective, can be contextualized as a *protection-oriented*, a *potential-oriented*, and a *policy-oriented* approach. It is *protection-oriented* when talking about qualifying workers for correct decision-making in cooperation with AI systems. It is *potential-oriented* when thinking about how to use AI systems for quality learning opportunities, e.g., in learning analytics. Finally, when reflecting on the advocacy tradition of empowering social subjects, an adult education perspective is *politically oriented*.

It is not surprising that from a HCI and psychology point of view, most perspectives by Wilkens et al. (2021) are relevant, as they touch technology and individual aspects at the same time. Interestingly, the *deficit-oriented* understanding and data *reliability-oriented* understanding perspectives are more related to HCI. So assisted tools that work through elaborated sensor technology are common to, e.g., gamified and explainable AI, which are fields of HCI.

In contrast, the *protection-oriented* understanding and *potential-oriented* understanding perspectives are closely related to psychology. Unrealistic user expectations questioning the "intelligence" of AI, In connection with loss of autonomy and trust, are prevailing psychological aspects of human-centricity for these two perspectives. Nevertheless, these areas can also be found in HCI and, depending on the degree of technical implementation, are likely to be assigned to HCI.

It is clear from Table 1 that the disciplines do not cover all of Wilkens et al. (2021) perspectives, but there are different emphases, with an imbalance existing for the first two perspectives. The article at hand shows that all disciplines are important for HCAI, with HFE functioning as an umbrella discipline for HCAI at work. This makes the call for an interdisciplinary view obvious. The following section will detail these views that are enriched by findings from a demonstrator.

6.2. Interdisciplinary views of human-centered AI

This section presents the different fields of action, the relevant factors of human-centered AI from different disciplines, and the interdisciplinary views of HCAI, including possible areas of collaboration. Adding insights from a demonstrator, the foundations for a synergistic human-AI symbiosis theory are revealed.

For *HFE*, human-centricity means involving humans in the design of work systems, e.g., processes or tools. The design is based on a comprehensive understanding of users, tasks, and work environments. Human-centered design aims at balanced skill and performance development as the basis of work productivity and health. *HFE* combines the human potential with technology. From a HFE perspective, human-centered AI emphasizes that the decision-making competence of humans and their intervention in technical systems in doubt are weighted higher than that of

TABLE 1 Analysis of human-centered AI perspectives for information science, adult education, human-computer interaction, and psychology according to Wilkens et al. (2021).

Discipline	Deficit-oriented (Perspective 1)	Data reliability-oriented (Perspective 2)	Protection-oriented (Perspective 3)	Potential-oriented (Perspective 4)	Political-oriented (Perspective 5)
Information science			x	x	x
Adult education			x	x	x
HCI	x	x			
Psychology			x	x	

AI machines. Information science and HCI provide details and key features of how to calibrate this interaction of humans and computers. A key factor lies in aspects of designing AI systems with regard to personality issues, search behavior, and information literacy of users.

From an *adult education* perspective, the success of human-centered AI lies in providing learning opportunities across all target groups affected by AI. These learning opportunities need to adapt to learners and their individual learning habits and learning needs in terms of content and methodology instead of one-fits-all solutions. Especially marginalized learners will have to be a focus. Information science can add to this perspective with its differentiated insights on information literacy, a concept that might be helpful when implementing HCAI solutions and, up-to-date, is far away from being profoundly treated in the field of adult education with, e.g., low-skilled adults. Furthermore, HFE perspectives can help in elaborating adult learning methodologies, as HFE provides clear perspectives on workplace learning and how it adds to effective learning.

Additionally, mutual reinforcement of humans and AI, which considers ethical principles during design, is essential and is actively considered in *information science*, e.g., in aspects of usability and interaction design. Especially the more profound understanding of sociotechnical information systems and how humans interact with digital environments are essential to information science.

Finally, for *HCI and psychology,* human-centered processes and products are the foundation of human-centricity. As a result, user experience design approaches and ongoing human performance evaluation are required. Additionally, human-centered AI products are emphasized by human control (Shneiderman, 2022, p. 9). These elements should be utilized to help tackle the six grand challenges of human-centered AI: human wellbeing, responsible design of AI, privacy aspects, AI-related design and evaluation frameworks, the role of government and independent oversight, and finally, HCAI interaction in general (Garibay et al., 2023). The overlap between HFE and HCI is visible where interaction design is considered from a perspective when knowledge and experience of usability and user experience are applied. This helps, among other things, to implement technology acceptance models and, thus, to build trust and support when the overlap between HFE and HCI is consistently implemented. In addition, HFE and psychology (and partly also HCI) meet in the field of human autonomy. Furthermore, the relationship between information science and HCI is exemplified in the interaction of any kind of information and the focus on human-centered systems, again considering usability and user experience. Recognition of the importance of information literacy should be considered an important link for HCAI in this regard.

In order to validate the findings compiled here, a demonstrator is used. The project "Connect & Collect: AI-based Cloud for Interdisciplinary Networked Research and Innovation for Future Work (CoCo)" (CoCo Website, 2023) promotes the transfer of knowledge between HFE research and operational practice in companies about artificial intelligence and is funded by the German Federal Ministry of Education and Research (BMBF). To support our theory development, the CoCo project serves as a demonstrator to illustrate the interrelation between the different disciplines. The purpose of the demonstrator is to reveal the necessity of each discipline, namely HFE, psychology, HCI, information science, and adult education, to create successful HCAI implementation at work and, therefore, to show the impact and potential in society. Participants are predominantly transdisciplinary actors in labor research from science, enterprises, unions, education, and intermediaries pursuing an innovative new approach or applying best practices and have joined forces in "Regional Competence Centers for Labor Research."

It can be derived from the demonstrator that while many companies are interested in implementing AI applications, they shy away from the research and investment effort involved in developing company-specific solutions. Instead, they aim to use proven AI applications. In this case, the importance of human-centered AI design-especially regarding learning facilitation and ethical-social compatibility-is not sufficiently applied (Pokorni et al., 2021). Undesirable consequences usually emerge only after a time delay and are rarely causally associated with AI use. An essential part of the work of the demonstrator is to increase the relevance of human-centered design of AI applications practically and systematically, which will also enhance the role of HFE experts.

At the same time, it is crucial to qualify the workers for the changes in the working society. The examples from the demonstrator show that, when implementing AI systems on a large scale, it is important to develop and establish a broad range of learning opportunities that can be upscaled to diverse target groups. Within the digital transformation, especially marginalized groups are at risk of getting lost in terms of workforce, labor markets, and in a democratic society. There is a need to focus on these marginalized groups when referring to human centeredness. Adult education providers and unions have an enormous role in this challenge because they have expertise in accessing the marginalized.

Human-centered AI is strongly related to HCI and psychology, albeit with different emphases. As HCI has evolved from a mainly

technical field to an interdisciplinary profession, the same can be expected for HCAI. Nevertheless, there are critical challenges to overcome, like explainability, trustworthiness, or unrealistic user expectations. However, technology acceptance models can be used to build trust to accept AI systems and make them more human-centric, thus keeping expectations in line. Ultimately, achieving an explainable AI contributes to the mutual collaboration and interaction of humans and AI. For ethical reasons, AI's instrumental character must always be considered. Furthermore, AI can be made more accessible to stakeholders by implementing gamification, which can increase stakeholder commitment to increased engagement with human-centered AI, especially in the workplace. Finally, the demonstrator acknowledges the importance of ethical concerns by utilizing the human-in-the-loop concept, thus increasing trust.

It can be concluded that the transfer of human-centered research results into practical application is supported by an interdisciplinary approach that combines different ideas, knowledge, and work methods. By displaying the interrelationships between the disciplines, this article reveals in the following section further directions for research as well as concepts and features a (future) theory of HCAI should entail.

6.3. Setting foundations for a human-centered interdisciplinary AI theory: synergistic human-AI symbiosis theory (SHAST)

In order to advance the research field regarding human-centered AI, a conceptualization of a Synergistic Human-AI Symbiosis Theory (SHAST) has been started. SHAST takes into account that the optimal deployment of artificial intelligence in the workplace needs the establishment of a symbiotic relationship between humans and AI systems, drawing upon the expertise of five distinct disciplines: human factors engineering (HFE), human-computer interaction (HCI), psychology, information science, and adult education. Based on the present findings, all five disciplines appear to be necessary to successfully implement HCAI in the workplace. SHAST envisions a future where AI systems and humans collaborate synergistically to achieve unprecedented levels of productivity and wellbeing. The five disciplines presented here are predestined to contribute to the future of HCAI, as they incorporate fundamental components of human centeredness and address important fields of action for its creation and implementation.

SHAST posits that AI should be designed to augment human capabilities, foster seamless interactions, and ensure ethical practices. HFE is the foundational pillar, advocating for AI systems that enhance human potential while preserving human autonomy, decision-making, and overall work performance. HCI, another elementary bedrock, focuses on user-centric design, creating seamless and intuitive interactions between humans and AI, fostering realistic user expectations, and minimizing friction in collaborating with AI technologies that intuitively adapt to user needs and preferences. Psychology's role in SHAST revolves around cultivating user trust, achieved through transparent AI design and explainable algorithms. Information science considers the pragmatic side of information, ensuring that humans can effectively and efficiently use information systems, for example, by increasing their information literacy. Finally, adult education plays a critical part in SHAST by cultivating digital literacy and ensuring that individuals possess the skills to navigate AI-powered environments, fostering learning opportunities for a workforce to engage with AI technologies for innovation and productivity effectively, minimizing disparities, and enabling broad participation.

TABLE 2 Framework with minimum fields of action for the disciplines.

Discipline	Minimum fields of action
HFE	• Balanced workload
	• Enhancing human capabilities
	• Human control
	• Social and ethical responsibility
Information science	• Pragmatics of information
	• Consider contexts and information behavior
	• Information literacy
Adult education	• Include low-skilled workers and workers' unions
	• Tailored learning and teaching in courses and the workplace
HCI and Psychology	• Technology acceptance and adoption
	• Human autonomy and control
	• Realistic user expectations
	• Explainable AI
	• Trust

SHAST proposes that human-AI symbiosis can be achieved through an interplay of these disciplines, resulting in AI technologies that empower individuals, enhance collaboration, and create a sustainable and equitable future. In fact, SHAST is based on a framework that is presented in Table 2. The outcomes from theory and practice show minimum fields of action for a successful HCAI implementation.

If fundamental aspects from the framework of the five disciplines are missing for the implementation and application of HCAI, then this will result in severe consequences and challenges (Stephanidis et al., 2019; Garibay et al., 2023; Xu et al., 2023). More precisely, it is postulated that when a discipline is not adequately considered, there may be a failure to comply with the minimum fields of action, and implementation may not be successful.

It remains open for discussion and empirical research on which disciplines would further be needed to support the development of a widely applicable theory of human-centered AI. The article aims to convincingly present the normative, theoretical, and methodological concepts from human factors and ergonomics (HFE), psychology, human-computer interaction, information science, and adult education and why they are considered critical building blocks for HCAI and SHAST.

7. Conclusions for human-centered AI at work from theory and practice

In the last section, the article summarizes how it contributes to the ongoing scientific discussion on HCAI. Conclusions are drawn for HCAI at work by leveraging insights from disciplines focusing mainly on individuals (psychology), technology (HCI), work (HFE), or work context as one research field amongst others (information science and adult learning). From the fundamental disciplinary considerations and the current experiences and observations from the demonstrator, theoretical implications are derived to bring HCAI in line with today's demands of workers and companies to reflect human centeredness when dealing with the complexity of information, data, and decisions. Furthermore, the article also highlights the relevant internal logic of the individual disciplines and reveals possible mutual complementarity. It mainly argues that, although human-centered AI is a popular concept across disciplines today (Capel and Brereton, 2023), successful HCAI and its design are in strong need for an interdisciplinary approach, as all disciplines conceptualize "their humans" differently in their views and methodological approaches.

Besides the differences in detail, our comparison of the different disciplinary approaches to conceptualizing HCAI and bringing it into practice has revealed important similarities. All presented disciplines have the human at their cores — the development of human-centered AI systems is therefore deeply connected to aspects considered central to humans and that cannot be substituted with machines, such as learning and constructing information. Technology, along with its AI systems, is considered a means for human development-therefore putting the human in a superior position, which also has to be guaranteed by certain regulations to prevent AI from overruling human decision-making. The disciplines also widely agree that human-centricity can only be achieved if struggles with the use of technology are minimized. In the end, it is always a human being in a company who will use technology. Depending on the degree of how much the AI design and AI implementation process reflects the users, including their skills, trust, and their work tasks — the users will cope or fail with technology within their workplace. Hence, raising the issue of humans practically coping or failing with technology use in the workplace as a common theoretical and practical problem of human centeredness might open scientific discourse between disciplines and practice in companies (Nowotny et al., 2001).

This article is based on collaborative scholarly inquiry and a review of discipline-specific literature on the phenomenon of HCAI and joint reflection against the background of the demonstrator used in our research to study HCAI. This led to the findings and arguments in this article, which are summarized in Table 2. The article combines perspectives from five disciplines-however, we are still in the first step, i.e., conceptualization, of a five-phase model for theory building in the applied sciences (Swanson and Chermack, 2013). The need to include further or remove disciplines will become obvious further down the road of the Theory-Research-Practice Cycle that is followed and will include operationalization, confirmation, application, and refinement (Swanson and Chermack, 2013) of the concept we have come up with. Hence, this research is not finished but at the beginning of understanding what might be critical for HCAI at work.

The first results of the conceptualization stage concerned with successful HCAI implementation at work and with the common focal points of the five disciplines are the founding pillars for a Synergistic Human-AI Symbiosis Theory (SHAST), which answers the research question of what is critical for HCAI at work. It has become apparent, however, that determining the critical aspects of HCAI at work, enabling a symbiotic relationship between humans and AI systems, and describing their interplay theoretically is a complex task. Whether the challenge does not rather require investigation of more minimum fields of action, more disciplinary backgrounds like, e.g., knowledge management, economics, process management, simulation theory, operation research, philosophy, history, sociology, and many more, or a transdisciplinary approach (Defila and Di Giulio, 2018) involving all stakeholders in the research, development, and implementation process of HCAI at work remains an open question. It has been argued that HCAI and human centeredness will benefit from considering many different scientific perspectives.[1] Furthermore, in today's world, disciplinary boundaries are fluid and increasingly blurred, or new disciplines are formed that encompass various aspects of other classical disciplines. We are aware of this, but due to a simplified discussion, we limit ourselves to a few disciplines and outline them, knowing that these boundaries are artificial.

The results from the demonstrator and our collaborative inquiry have shown that the disciplines covered in this article have many similar (and some different) concepts and methods in their portfolios that are most likely critical for successfully implementing HCAI at work. However, a broader interdisciplinary discussion and research are needed for a complete view of the topic.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

AM: Conceptualization, Writing—original draft, Writing—review & editing, Funding acquisition, Supervision, Methodology, Project administration, Validation, Visualization. CB-S: Conceptualization, Writing—original draft, Writing—review & editing, Methodology, Project administration, Validation, Visualization. MB: Conceptualization, Writing—original draft, Writing—review & editing, Methodology, Project administration, Validation, Visualization. IP: Conceptualization, Writing—original draft, Writing—review & editing, Methodology, Project administration, Validation, Visualization.

1 https://www.frontiersin.org/research-topics/50257/human-centered-ai-at-work-common-ground-in-theories-and-methods

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The CoCo project is funded by the Federal Ministry of Education and Research-grant numbers 02L19C000, 02L19C001, and 02L19C002.

Acknowledgments

We are grateful for the valuing and precise feedback of the reviewers. It helped us to improve the present article and to elaborate on its line of argument.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

ACRL (2016). *Framework for Information Literacy for Higher Education*. Chicago, IL: Association of College and Research Libraries.

Alheit, P. (2021). Biographicity as "mental grammar" of postmodern life. *Eur. J. Res. Educ. Learn. Adults* 12, 81–94. doi: 10.3384/rela.2000-7426.ojs1845

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., et al. (2019). "Guidelines for human-AI interaction," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems CHI '19* (New York, NY: ACM), 1–13.

Anandiadou, K., and Claro, M. (2009). *21st Century Skills and Competences for New Millennium Learners in OECD Countries, OECD Education Working Papers, No. 41*. Washington, DC: OECD Publishing.

Anderson, T., and Rivera-Vargas, P. (2020). A critical look at educational technology from a distance education perspective. *Digital Educ. Rev.* 37, 208–229. doi: 10.1344/der.2020.37.208-229

Arbeiter, J. (2023). Acting between professional rules and financial resources in the field of work-oriented basic education. *Zeitschrift für Weiterbildungsforschung* 46, 87–102. doi: 10.1007/s40955-023-00240-2

Autor:innengruppe Bildungsberichterstattung (2022). *Bildung in Deutschland 2022. Ein indikatorengestützter Bericht mit einer Analyse zum Bildungspersonal*. Bielefeld: wbv Media.

Bagozzi, R. P. (2007). The legacy of the technology acceptance model and a proposal for a paradigm shift. *J. Assoc. Inf. Syst.* 8, 244–254. doi: 10.17705/1jais.00122

Bates, M. J. (2017). "Information behavior," in *Encyclopedia of Library and Information Sciences. 4th Edn*, eds. J. McDonald and M. Levine-Clark (Boca Raton: CRC Press), 2074–2085.

Bauer, J., Festner, D., Gruber, H., Harteis, C., and Heid, H. (2004). The effects of epistemological beliefs on workplace learning. *J. Workpl. Learn.* 16, 284–292. doi: 10.1108/13665620410545561

Baumeister, H., Garatva, P., Pryss, R., Ropinski, T., and Montag, C. (2023). Digitale Phänotypisierung in der Psychologie - ein Quantensprung in der psychologischen Forschung? *Psychol. Rundsch*. 74, 89–106. doi: 10.1026/0033-3042/a000609

Baxter, G., and Sommerville, I. (2011). Socio-technical systems: from design methods to systems engineering. *Interacting Comput.* 23, 4–17. doi: 10.1016/j.intcom.2010.07.003

Bender, E., and Kolle, A. (2020). "Climbing towards NLU–on meaning, form, and understanding in the age of data," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Seattle, WA: Association for Computational Linguistics, 5185–5198.

Bennett, D., Metatla, O., Roudaut, A., and Mekler, E. D. (2023). "How does HCI understand human agency and autonomy?" in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems CHI '23* (New York, NY: Association for Computing Machinery), 1–18.

Bermúdez, J. P., Nyrup, R., Deterding, S., Mougenot, C., Moradbakhti, L., You, F., et al. (2023). "What is a subliminal technique? An ethical perspective on AI-driven influence," in *Proceedings of the IEEE ETHICS-2023 Conference* (West Lafayette, IN: IEEE), 1–9.

BMAS and BMBF (2021). *Nationale Weiterbildungsstrategie*. Available online at: https://www.bmas.de/SharedDocs/Downloads/DE/Aus-Weiterbildung/nws-fortfuehrung-und-weiterentwicklung.pdf?__blob=publicationFile&v=3 (accessed June 29, 2023).

BMBF (2022). *Weiterbildungsverhalten in Deutschland 2020. Ergebnisse des Adult Education Survey—AES-Trendbericht*. Bonn: BMBF.

Böhle, F. (2009). "Weder Rationale Reflexion noch Präreflexive Praktik," in *Erfahrungsgeleitet-Subjektivierendes Handeln, Handeln unter Unsicherheit*, eds. F. Böhle and M. Weihrich (Wiesbaden: Springer VS), 203–230.

Bostrom, N., and Yudkowsky, E. (2013). "The ethics of artificial intelligence," in *Cambridge Handbook of Artificial Intelligence*, eds. K. Frankish, M. Willian, and W. M. Ramsey (Cambridge: University Press), 316–334.

Bräuer, P., and Mazarakis, A. (2022a). "Alexa, can we design gamification without a screen?"-Implementing cooperative and competitive audio-gamification for intelligent virtual assistants. *Comput. Hum. Behav.* 135, 107362. doi: 10.1016/j.chb.2022.107362

Bräuer, P., and Mazarakis, A. (2022b). How to design audio-gamification for language learning with amazon Alexa? a long-term field experiment. *Int. J. Hum. Comput. Int.* 12, 1–18. doi: 10.1080/10447318.2022.2160228

Braun, M. (2017). "Arbeit 4.0: Der gesunde Mensch in der digitalisierten Arbeitswelt," in *Handbuch der Arbeitsmedizin*, eds D. Nowak, and S. Letzel (Landsberg: Ecomed), 1–24.

Brynjolfsson, E., and McAfee, A. (2012). *Race Against the Machine: How the Digital Revolution is Accelerating Innovation, Driving Productivity, and Irreversibly Transforming Employment and the Economy*. Lexington MA: Digital Frontier Press.

Bülchmann, O. (2020). Künstliche Intelligenz und Ethik–ein ungleiches Paar? *Wirtschaftsinformatik Manage.* 12, 206–215. doi: 10.1365/s35764-020-00256-0

Capel, T., and Brereton, M. (2023). "What is human-centered about human-centered AI? A map of the research landscape," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems CHI '23* (New York, NY: Association for Computing Machinery), 1–23. doi: 10.1145/3544548.3580959

Case, D. O., and Given, L. M. (2016). *Looking for Information: A Survey of Research on Information Seeking, Needs, and Behavior, 4th Edn*. Bingley: Emerald Group Publishing.

Chai, C. S., Lin, P. Y., Jong, M. S. Y., Dai, Y., Chiu, T. K., and Huang, B. (2020). "Factors influencing students' behavioral intention to continue artificial intelligence learning," in *2020 International Symposium on Educational Technology (ISET)*. Piscataway, NJ: IEEE.

Choung, H., David, P., and Ross, A. (2023). Trust in AI and its role in the acceptance of AI technologies. *Int. J. Hum. Comput. Int.* 39, 1727–1739. doi: 10.1080/10447318.2022.2050543

Clemmensen, T. (2006). Whatever happened to the psychology of human-computer interaction? *Inf. Technol. People* 19, 121–151. doi: 10.1108/09593840610673793

CoCo Website (2023). *Connect & Collect: KI-Gestützte Cloud für die Interdisziplinäre vernetzte Forschung und Innovation für die Zukunftsarbeit*. Available online at: https://www.coco-projekt.de/ (accessed February 4, 2023).

Council of the European Union (2021). *Council Resolution on a New European Agenda for Adult Learning 2021-2030 (2021/C 504/02)*.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q.* 21, 319–340. doi: 10.2307/249008

Defila, R., and Di Giulio, A. (2018). *Transdisziplinär und Transformativ Forschen*. Wiesbaden: Springer Fachmedien.

Dehnbostel, P. (2022). *Betriebliche Bildungsarbeit. Kompetenzbasierte Berufs-und Weiterbildung in digitalen Zeiten*. Baltmannsweiler: Schneider Verlag Hohengehren.

Del Giudice, M., Scuotto, V., Orlando, B., and Mustilli, M. (2023). Toward the human–centered approach. a revised model of individual acceptance of AI. *Hum. Resour. Manag. Rev.* 33, 100856. doi: 10.1016/j.hrmr.2021.100856

Deterding, S., Dixon, D., Khaled, R., and Nacke, L. (2011). "From Game Design Elements to Gamefulness: Defining 'Gamification.'," in *Proceedings of the 15th international Academic MindTrek Conference: Envisioning Future Media Environments* (London: ACM Press), 9–15.

Dollhausen, K., and Lattke, S. (2020). "Organisation und Organisationsformen wissenschaftlicher Weiterbildung," in *Handbuch wissenschaftliche Weiterbildung*, eds. W. Jütte and M. Rohs (Wiesbaden: Springer VS), 99–121.

Dyckhoff, H. (2006). *Produktionstheorie, 5th Edn*. Berlin: Springer.

Elbeshausen, S. (2023). "C 10 Modellierung von Benutzer*innen, Kontextualisierung, Personalisierung," in *Grundlagen der Informationswissenschaft. 7th Edn*, eds. R. Kuhlen, D. Lewandowski, W. Semar and C. Womser-Hacker (Berlin: De Gruyter Saur), 467–476.

European Union (2021). *Council Resolution on a New European Agenda for Adult Learning 2021-2030 (2021/C 504/02). Offical Journal of the European Union.* Available online at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32021G1214(01) (accessed August 28, 2023).

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., et al. (2018). AI4People—An ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Mind Mach* 28, 689–707. doi: 10.1007/s11023-018-9482-5

Garibay, O. O., Winslow, B., Andolina, S., Antona, M., Bodenschatz, A., Coursaris, C., et al. (2023). Six human-centered artificial intelligence grand challenges. *Int. J. Hum. Comput. Int.* 39, 391–437. doi: 10.1080/10447318.2022.2153320

Gericke, E. (2017). Why returning to VET? – Results of a qualitative comparative. *study about English and German Mechatronics. International Journal for Res. Vocat. Educ. Train.* 4, 206–225. doi: 10.13152/IJRVET.4.3.2

Golmann, R., Hagmann, D., and Loewenstein, G. (2017). Information avoidance. *J. Econ. Lit.* 55, 96–135. doi: 10.1257/jel.20151245

Goppold, M., and Frenz, M. (2020). Lernen im Prozess der Arbeit: Entwicklung, Umsetzung und Evaluation einer Weiterbildung zur manuellen Montage unter Einsatz von Autorensystemen. *Zeitschrift für Arbeitswissenschaft* 74, 100–116. doi: 10.1007/s41449-020-00202-1

Greifeneder, E., and Schlebbe, K. (2023). "D 1 Information Behaviour. in Grundlagen der Informationswissenschaft," in *Grundlagen der Informationswissenschaft. 7th edn.*, eds. R. Kuhlen, D. Lewandowski, W. Semar and C. Womser-Hacker (Berlin, Boston: De Gruyter Saur), 497–510.

Griesbaum, J. (2023). "D 8 Informationskompetenz," in *Grundlagen der Informationswissenschaft, 7th Edn.*, eds. R. Kuhlen, D. Lewandowski, W. Semar, and C. Womser-Hacker (Berlin: De Gruyter Saur), 581–594.

Grote, G., Ryser, C., Wäfler, T., Windischer, A., and Weik, S. (2000). Kompass–a method for complementary function allocation in automated work systems. *Int. J. Hum. Comput. Stu.* 52, 267–287. doi: 10.1006/ijhc.1999.0289

Gust von Loh, S. (2008). Wissensmanagement und Informationsbedarfsanalyse in kleinen und mittleren Unternehmen: Teil 2: Wissensmanagement in KMU. *Inf.-Wissenschaft Praxis* 59, 127–135.

Habayeb, A. (2022). *Explainable AI Isn't Enough; We Need Understandable AI. Techopedia.* Available online at: https://www.techopedia.com/explainable-ai-isnt-enough-we-need-understandable-ai/2/34671 (accessed June 24, 2023).

Hacker, W. (2018). *Menschengerechtes Arbeiten in der Digitalisierten Welt*. Zürich: VDF.

Haghighat, P., Nguyen, T., Valizadeh, M., Arvan, M., Parde, N., Kim, M., et al. (2023). Effects of an intelligent virtual assistant on office task performance and workload in a noisy environment. *Appl. Ergon.* 109, 103969. doi: 10.1016/j.apergo.2023.103969

Harteis, C. (2022). "Research on workplace learning in times of digitalisation," in *Research Approaches on Workplace Learning: Insights from a Growing Field*, eds. C. Harteis, D. Gijbels, and E. Kyndt (Cham: Springer International Publishing) 415–428. doi: 10.1007/978-3-030-89582-2

Hartel, J. (2019). *Turn, Turn, Turn, Proceedings of the Tenth International Conference on Conceptions of Library and Information Science. Ljubljana (Slovenia), June 16th-19th.* https://informationr.net/ir/24-4/colis/colis1901.html (accessed June 29, 2023).

Henrichs, N. (2014). "Sozialisation der information. Zum aufgabenspektrum der informationswissenschaft," in *Menschsein im Informationszeitalter. Informationswissenschaft mit Leidenschaft und Missionarischem Eifer,* ed. N. Henrichs (Glückstadt: Verlag Werner Hülsbusch).

Höök, K. (2000). Steps to take before intelligent user interfaces become real. *Interact. Comput.* 12, 409–426. doi: 10.1016/S0953-5438(99)00006-5

Huchler, N. (2022). Komplementäre Arbeitsgestaltung. Grundrisse eines Konzepts zur Humanisierung der Arbeit mit KI. *Zeitschrift für Arbeitswissenschaft* 76, 158–175. doi: 10.1007/s41449-022-00319-5

Huchler, N., Adolph, L., Andre, E., Bauer, W., Reißner, N., Müller, N., et al. (2020). *Kriterien für die Mensch–Maschine-Interaktion bei KI. Ansätze für die menschengerechte Gestaltung in der Arbeitswelt–Whitepaper aus der Plattform Lernende Systeme.* München: ACATECH.

IEA (2000). *What is Ergonomics?* Available online at: https://iea.cc/about/what-is-ergonomics/ (accessed July 11, 2023).

ILO and IEA (2021). *Principles and Guidelines for (HFE) Design and Management of Work Systems*. Geneva: Joint Document by ILO and IEA.

Infratest dimap (2023). *KI ist eher Risiko als Chance, nicht jedoch für Junge und Bessergebildete. Umfragen Anal.-Infratest Dimap*. Available online at: https://www.infratest-dimap.de/umfragen-analysen/bundesweit/umfragen/aktuell/kritischer-ausblick-auf-deutsch-tuerkische-beziehungen-bei-wiederwahl-erdogans/ (accessed May 12, 2023).

Ingwersen, P., and Järvelin, K. (2005). *The Turn: Integration of Information Seeking and Retrieval in Context.* Cham: Springer Science and Business Media.

Ipsos (2022). *Opinions About AI Vary Depending on Countries' Level of Economic Development. Ipsos*. Available online at: https://www.ipsos.com/en/global-opinions-about-ai-january-2022 (accessed May 13 2023).

ISO 6385 (2016). *Ergonomics Principles in the Design of Work Systems*.

ISO 9241-110 (2020). *Ergonomics of Human-System Interaction - Part 110: Interaction Principles*.

ISO 9241-2 (1992). *Ergonomic Requirements for Office Work With Visual Display Terminals (VDTs) - Part 2: Guidance on Task Requirements*.

ISO 9241-210 (2019). *Ergonomics of Human-System Interaction - Part 210: Human-Centred Design for Interactive Systems*.

Jetter, H. C. (2023). "D 3 Mensch-Computer-Interaktion, Usability und User Experience," in *Grundlagen der Informationswissenschaft, 7th Edn,* eds. R. Kuhlen, D. Lewandowski, W. Semar, and C. Womser-Hacker (Berlin, Boston: De Gruyter Saur), 525–534.

Joint Research Center (JRC) (2022). *DigComp 2.2: The Digital Competence Framework for Citizens-With New Examples of Knowledge, Skills and Attitudes*. EU Publications. doi: 10.2760/115376

Jones, W. (2010). No knowledge but through information. *First Monday* 15, 3602. doi: 10.5210/fm.v15i9.3062

Kao, W.-K., and Huang, Y.-S. (2023). Service robots in full-and limited-service restaurants: extending technology acceptance model. *J. Hosp. Tour. Manag.* 54, 10–21. doi: 10.1016/j.jhtm.2022.11.006

Käpplinger, B., and Lichte, N. (2020). The lockdown of physical co-operation touches the heart of adult education: a delphi study on immediate and expected effects of COVID-19. *Int. Rev. Educ.* 66, 777–795. doi: 10.1007/s11159-020-09871-w

Karwowski, W., Szopa, A., and Soares, M. (2021). *Handbook of Standards and Guidelines in Human Factors and Ergonomics*. Boca Raton, WA: CRC Press.

Kosinski, M., Bachrach, Y., Kohli, P., Stillwell, D., and Graepel, T. (2014). Manifestations of user personality in website choice and behaviour on online social networks. *Mach. Learn.* 95, 357–380. doi: 10.1007/s10994-013-5415-y

Kraus, K. (2008). "Does employability put the german 'vocational order' at risk? An analysis from the perspective of earning oriented pedagogy," in *Work, Education and Employability*, eds. P. Gonon, K. Kraus, J. Oelkers, and S. Stolz Dernbach (Bern: Peter Lang), 55–81.

Krishnan, C., Gupta, A., Gupta, A., and Singh, G. (2022). "Impact of artificial intelligence-based chatbots on customer engagement and business growth," in *Deep Learning for Social Media Data Analytics* Studies in Big Data., eds. T.-P. Hong, L. Serrano-Estrada, A. Saxena, and A. Biswas (Cham: Springer International Publishing), 195–210.

Kuhlen, R. (2004). *Information Grundlagen der praktischen Information und Dokumentation, 5th Edn*. München: Saur, 3–20.

Kurni, M., Mohammed, M. S., and Srinivasa, K. G. (2023). "AI-enabled gamification in education," in *A Beginner's Guide to Introduce Artificial Intelligence in Teaching and Learning*, eds. M. Kurni, M. S. Mohammed, and S. K G (Cham: Springer), 105–114.

Laato, S., Tiainen, M., Najmul Islam, A. K. M., and Mäntymäki, M. (2022). How to explain AI systems to end users: a systematic literature review and research agenda. *Int. Res.* 32, 1–31. doi: 10.1108/INTR-08-2021-0600

Lammers, A., Lukowski, F., and Weis, K. (2022). The relationship between works councils and firms' further training provision in times of technological change. *Br. J. Ind. Relat.* 61, 392–424. doi: 10.1111/bjir.12710

Landgrebe, J., and Smith, B. (2022). *Why Machines Will Never Rule the World: Artificial Intelligence Without Fear*. Abingdon: Routledge.

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., et al. (2021). What do we want from explainable artificial intelligence (XAI)?–A stakeholder perspective on Xai and a conceptual model guiding interdisciplinary xai research. *Artif. Int.* 296, 103473. doi: 10.1016/j.artint.2021.103473

Leichtmann, B., Hinterreiter, A., Humer, C., Streit, M., and Mara, M. (2023a). Explainable artificial intelligence improves human decision-making: results from a mushroom picking experiment at a public art festival. *Int. J. Hum. Comput. Int.* 12, 1–18. doi: 10.1080/10447318.2023.2221605

Leichtmann, B., Humer, C., Hinterreiter, A., Streit, M., and Mara, M. (2023b). Effects of explainable artificial intelligence on trust and human behavior in a high-risk decision task. *Comput. Hum. Behav.* 139, 107539. doi: 10.1016/j.chb.2022.107539

Lenzen, M. (2019). *Künstliche Intelligenz: Was sie kann & was uns erwartet*. München: Beck.

Lewandowski, D., and Womser-Hacker, C. (2023). "D 6 information seeking behaviour," in *Grundlagen der Informationswissenschaft, 7th Edn*, eds. R. Kuhlen, D. Lewandowski, W. Semar, and C. Womser-Hacker (Berlin: De Gruyter Saur), 553–566.

Lloyd, A. (2013). "Building information resilient workers: the critical ground of workplace information literacy. What have we learnt?" in *Worldwide Commonalities and Challenges in Information Literacy Research and Practice: European Conference on Information Literacy, ECIL 2013,* Istanbul: Springer International Publishing, 219–228.

Luger, E., and Sellen, A. (2016). ""Like having a really bad PA": the gulf between user expectation and experience of conversational agents," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems CHI '16*. New York, NY: Association for Computing Machinery (ACM), 5286–5297.

Marsden, N., and Haag, M. (2016). "Stereotypes and politics: reflections on personas," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems CHI '16*. New York, NY: Association for Computing Machinery (ACM), 4017–4031.

Mazarakis, A. (2013). Like diamonds in the sky–how feedback can boost the amount of available data for learning analytics. *Int. J. Technol. Enhanc. Learn.* 5, 107–116. doi: 10.1504/IJTEL.2013.059083

Mazarakis, A. (2021). Gamification reloaded: current and future trends in gamification science. *i-com* 20, 279–294. doi: 10.1515/icom-2021-0025

McCarthy, J. (2007). *What is Artificial Intelligence? Stanford University*. Available online at: https://www-formal.stanford.edu/jmc/whatisai.pdf (accessed August 29, 2023).

McLean, S., Read, G. J., Thompson, J., Baber, C., Stanton, N. A., and Salmon, P. M. (2021). The risks associated with artificial general intelligence: a systematic review. *J. Exp. Theor. Artif. Int.* 24, 1–15. doi: 10.1080/0952813X.2021.1964003

Merriam, S. B., and Bierema, L. L. (2013). *Adult Learning: Linking Theory and Practice*. London: John Wiley & Sons.

Middleton, L., Hall, H., Muir, L., and Raeside, R. (2018). The interaction between people, information and innovation: information literacy to underpin innovative work behaviour in a Finnish organization. *Proc. Assoc. Inf. Sci. Technol.* 55, 367–376. doi: 10.1002/pra2.2018.14505501040

Miller, T. (2019). Explanation in artificial intelligence: insights from the social sciences. *Artif. Int.* 267, 1–38. doi: 10.1016/j.artint.2018.07.007

Ng, D. T. K., Leung, J. K. L., Chu, K. W. S., and Qiao, M. S. (2021a). AI literacy: definition, teaching, evaluation and ethical issues. *Proc. Assoc. Inf. Sci. Technol.* 58, 504–509. doi: 10.1002/pra2.487

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., and Qiao, M. S. (2021b). Conceptualizing AI literacy: an exploratory review. *Comput. Educ. Artif. Int.* 2, 100041. doi: 10.1016/j.caeai.2021.100041

Nowotny, H., Scott, P. B., and Gibbons, M. (2001). *Re-Thinking Science. Knowledge and the Public in an Age of Uncertainty*. Cambridge, MA: Polity.

OECD (2023). *Equity and Inclusion in Education: Finding Strength through Diversity*. Paris: OECD Publishing.

OECD. (2019). *Artificial Intelligence in Society*. Paris: OECD Publishing.

Pirolli, P., and Card, S. (1999). Information foraging. *Psychol. Rev.* 106, 643–675. doi: 10.1037/0033-295X.106.4.643

Pokorni, B., Braun, M., and Knecht, C. (2021). *Menschzentrierte KI-Anwendungen in der Produktion. Praxis-erfahrungen und Leitfaden zu Betrieblichen Einführungsstrategien. Projektbericht im Fortschrittszentrum "Lernende Systeme"*. Stuttgart: Fraunhofer IAO.

Raftopoulos, M., and Hamari, J. (2023). "Artificial intelligence in the workplace: implementation challenges and opportunities," in *Proceedings of the Americas Conference on Information Systems (AMCIS) 2023*. Cancún: AMCIS.

Rammert, W. (2009). "Hybride Handlungsträgerschaft: Ein soziotechnisches Modell verteilten Handelns," in *Intelligente Objekte*, eds. O. Herzog and T. Schildhauer (Berlin: Springer), 23–33.

Rauch, W. (1988). *Was ist Informationswissenschaft?* Graz: Kienreich.

Reddy, M. C., and Jansen, B. J. (2008). A model for understanding collaborative information behavior in context: a study of two healthcare teams. *Inf. Proc. Manage.* 44, 256–273. doi: 10.1016/j.ipm.2006.12.010

Riege, A. (2005). Three-dozen knowledge-sharing barriers managers must consider. *J. Knowl. Manag.* 9, 18–35. doi: 10.1108/13673270510602746

Rohs, M., and Ganz, M. (2015). MOOCs and the claim of education for all: a disillusion by empirical data. *Int. Rev. Res. Open Distrib. Learn.* 16, 1–19. doi: 10.19173/irrodl.v16i6.2033

Schlick, C., Bruder, R., and Luczak, H. (2018). *Arbeitswissenschaft, 4th Edn*. Berlin, Heidelberg: Springer Vieweg.

Schnabel, D. (2017). *Kompetenzen für die Arbeitswelt Von Heute und Morgen: 21st Century Skills and Beyond*. Available online at: https://hochschulforumdigitalisierung.de/de/blog/kompetenzen-21st-century-skills#:$\sim$:text=Das%2021st%20Century%20Skills%20Modell%20deckt%20vier%20Kompetenzfelder,Kritisches%20Denken%204%20Flexibilit%C3%A4t%2C%20Ambiguit%C3%A4tstoleranz%2C%20Eigenmotivation%2C%20Selbst%C3%A4ndiges%20Arbeiten (accessed June 29, 2023).

Schön, D. (1983). *The Reflective Practitioner: How Professionals Think in Action*. New York, NY: Basic Books.

Shannon, C. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mob. Comput. Commun. Rev.* 5, 3–55. doi: 10.1145/584091.584093

Shneiderman, B. (2020). Human-centered artificial intelligence: three fresh ideas. *AIS Trans. Hum.-Comput. Interact.* 12, 109–124. doi: 10.17705/1thci.00131

Shneiderman, B. (2022). *Human-Centered AI*. New York, NY: Oxford University Press.

Sohn, K., and Kwon, O. (2020). Technology acceptance theories and factors influencing artificial intelligence-based intelligent products. *Telemat. Inform.* 47, 101324. doi: 10.1016/j.tele.2019.101324

Spitzley, H. (1980). *Wissenschaftliche Betriebsführung, REFA-Methodenlehre und Neuorientierung der Arbeitswissenschaft*. Frankfurt: Bund.

Stephanidis, C., Salvendy, G., Antona, M., Chen, J. Y. C., Dong, J., Duffy, V. G., et al. (2019). Seven HCI grand challenges. *Int. J. Hum. Comput. Int.* 35, 1229–1269. doi: 10.1080/10447318.2019.1619259

Stock, W., and Stock, M. (2015). *Handbook of Information Science*. Berlin: De Gruyter Saur.

Swanson, R. A., and Chermack, T. J. (2013). *Theory Building in Applied Disciplines*. San Francisco, CA: Berrett-Koehler Publishers.

Tan, D. Y., and Cheah, C. W. (2021). Developing a gamified AI-enabled online learning application to improve students' perception of university physics. *Comput. Educ. Artif. Intell.* 2, 100032. doi: 10.1016/j.caeai.2021.100032

Tang, P. M., Koopman, J., Mai, K. M., De Cremer, D., Zhang, J. H., Reynders, P., et al. (2023). No person is an island: unpacking the work and after-work intelligence. *J. Appl.* consequences of interacting with artificial *Psychol.* 14, 1–24. doi: 10.1037/apl0001103

Tocchetti, A., Corti, L., Brambilla, M., and Celino, I. (2022). EXP-crowd: a gamified crowdsourcing framework for explainability. *Front. Artif. Intell.* 5, 826499. doi: 10.3389/frai.2022.826499

Touretzky, D., Gardner-McCune, C., Martin, F., and Seehorn, D. (2019). Envisioning AI for K-12: what should every child know about AI? *Proc. AAAI Conf. Artif. Int.* 33, 9795–9799. doi: 10.1609/aaai.v33i01.3301979

Travis, T. A. (2017). From the classroom to the boardroom: the impact of information literacy instruction on workplace research skills. *Educ. Lib.* 34, 19–31. doi: 10.26443/el.v34i2.308

Ulich, E. (2011). *Arbeitspsychologie. 7th Edn*. Zürich: VDF Hochschulverlag.

UNESCO (2013). *Global media and Information Literacy (MIL) Assessment Framework: Country Readiness and Competencies*. Paris: UNESCO.

UNESCO (2015). *Global Network of Learning Cities*. Guiding Document. Hamburg: UNESCO Institute for Lifelong Learning.

UNESCO (2019). *4th Global Report on Adult Learning and Education*. Hamburg: UNESCO Institute for Lifelong Learning.

Venkatesh, V., Morris, M. G., Davis, G. B., and Davis, F. D. (2003). User acceptance of information technology: toward a unified view. *MIS Q.* 12, 425–478. doi: 10.2307/30036540

Wahde, M., and Virgolin, M. (2023). DAISY: an implementation of five core principles for transparent and accountable conversational AI. *Int. J. Hum. Comput. Int.* 39, 1856–1873. doi: 10.1080/10447318.2022.2081762

Watkins, B. J., and Tisdell, E. J. (2006). Negotiating the labyrinth from margin to center: adult degree program administrators as program planners within Higher Education Institutions. *Adult Educ. Q.* 56, 134–159. doi: 10.1177/0741713605283433

Wersig, G. (1974). *Information-Kommunikation-Dokumentation*. Darmstadt: Wissenschaftliche Buchgesellschaft.

Westphal, M., Vössing, M., Satzger, G., Yom-Tov, G. B., and Rafaeli, A. (2023). Decision control and explanations in human-AI collaboration:

improving user perceptions and compliance. *Comput. Hum. Behav*. 144, 107714. doi: 10.1016/j.chb.2023.107714

Whorf, B. L. (1963). *Sprache, Denken, Wirklichkeit. Beiträge zur Metalinguistik und Sprachphilosophie.* Hamburg: Rowohlt.

Wilkens, U., Reyes, C. C., Treude, T., and Kluge, A. (2021). *Understandings and Perspectives of Human-Centered AI-a Transdisciplinary Literature Review. In Bericht zum 67. Arbeitswissenschaftlichen Kongress.* Dortmund: GfA-Press.

Wilkens, U., Süße, T., and Voigt, B.-F. (2014). "Umgang mit Paradoxien von Industrie 4.0–Die Bedeutung reflexiven Arbeitshandelns," in *Industrie 4.0–Wie intelligente Vernetzung und kognitive Systeme unsere Arbeit verändern*, eds. W. Kersten, H. Koller, and H. Lödding (Berlin: GITO), 199–210.

Wilson, J. R. (2014). Fundamentals of systems of ergonomics/human factors. *Appl. Ergon.* 45, 5–13. doi: 10.1016/j.apergo.2013.03.021

Wilson, T. D. (1999). Models in information behaviour research. *J. Document.* 55, 249–270. doi: 10.1108/EUM0000000007145

Wilson, T. D. (2000). Human information behavior. *Inf. Sci.* 3, 49–55. doi: 10.28945/576

Xu, W., Dainoff, M. J., Ge, L., and Gao, Z. (2023). Transitioning to human interaction with ai systems: new challenges and opportunities for HCI professionals to enable human-centered AI. *Int. J. Hum. Comput. Int*. 39, 494–518. doi: 10.1080/10447318.2022.2041900

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. San Diego, CA: Hafner Publishing Company.

frontiers | Frontiers in Artificial Intelligence

TYPE Systematic Review
PUBLISHED 01 November 2023
DOI 10.3389/frai.2023.1272159

OPEN ACCESS

EDITED BY
Assunta Di Vaio,
University of Naples Parthenope, Italy

REVIEWED BY
Tobias Ley,
Tallinn University, Estonia
Victor Lo,
Fidelity Investments, United States
Paola Briganti,
University of Naples Parthenope, Italy

*CORRESPONDENCE
Uta Wilkens
✉ uta.wilkens@ruhr-uni-bochum.de

RECEIVED 03 August 2023
ACCEPTED 29 September 2023
PUBLISHED 01 November 2023

CITATION
Wilkens U, Lupp D and Langholf V (2023) Configurations of human-centered AI at work: seven actor-structure engagements in organizations.
Front. Artif. Intell. 6:1272159.
doi: 10.3389/frai.2023.1272159

Configurations of human-centered AI at work: seven actor-structure engagements in organizations

Uta Wilkens*, Daniel Lupp and Valentin Langholf

Institute of Work Science, Ruhr University Bochum, Bochum, Germany

Purpose: The discourse on the human-centricity of AI at work needs contextualization. The aim of this study is to distinguish prevalent criteria of human-centricity for AI applications in the scientific discourse and to relate them to the work contexts for which they are specifically intended. This leads to configurations of actor-structure engagements that foster human-centricity in the workplace.

Theoretical foundation: The study applies configurational theory to sociotechnical systems' analysis of work settings. The assumption is that different approaches to promote human-centricity coexist, depending on the stakeholders responsible for their application.

Method: The exploration of criteria indicating human-centricity and their synthesis into configurations is based on a cross-disciplinary literature review following a systematic search strategy and a deductive-inductive qualitative content analysis of 101 research articles.

Results: The article outlines eight criteria of human-centricity, two of which face challenges of human-centered technology development (trustworthiness and explainability), three challenges of human-centered employee development (prevention of job loss, health, and human agency and augmentation), and three challenges of human-centered organizational development (compensation of systems' weaknesses, integration of user-domain knowledge, accountability, and safety culture). The configurational theory allows contextualization of these criteria from a higher-order perspective and leads to seven configurations of actor-structure engagements in terms of engagement for (1) data and technostructure, (2) operational process optimization, (3) operators' employment, (4) employees' wellbeing, (5) proficiency, (6) accountability, and (7) interactive cross-domain design. Each has one criterion of human-centricity in the foreground. Trustworthiness does not build its own configuration but is proposed to be a necessary condition in all seven configurations.

Discussion: The article contextualizes the overall debate on human-centricity and allows us to specify stakeholder-related engagements and how these complement each other. This is of high value for practitioners bringing human-centricity to the workplace and allows them to compare which criteria are considered in transnational declarations, international norms and standards, or company guidelines.

KEYWORDS

human-centered, artificial intelligence, AI, work, sociotechnical system, configurational theory, stakeholder

1. Introduction

Human-centered and responsible artificial intelligence (AI) applications are of key concern in current national and trans-national proposals for declarations and regulations, such as the US Blueprint AI Bill of Rights or the EU AI Act, of norming initiatives of the International Organization for Standardization (ISO), and of company guidelines, e.g., the Microsoft Responsible AI declaration or SAP's Guiding Principles for AI. At the same time, there is an academic-driven research debate to which different research communities contribute. This article sheds light on the criteria of human-centricity and how they are considered in academic publications. Whether and how they are treated in political declarations and industry norms will be part of the discussion.

Scholars elaborate on the meaning of human-centricity either of AI as a technology (Zhu et al., 2018; Ploug and Holm, 2020; How et al., 2020b), AI applications related to the work context (Jarrahi, 2018; Wilson and Daugherty, 2018; Gu et al., 2021), or job characteristics of work contexts in which AI applications are implemented (Romero et al., 2016; Kluge et al., 2021; Parker and Grote, 2022). Systematic overviews on these criteria show that contributing researchers are from a wide range of disciplines and include certain use fields such as healthcare, manufacturing, education, or administration, as well as the work processes of software development itself (Wilkens et al., 2021a). The thematic foci vary depending on the discipline and field of use. While researchers from the human-computer interaction (HCI) community describe human-centered AI as an issue of AIs' trustworthiness and related safety culture (Shneiderman, 2022) and thus combine technological characteristics with organizational characteristics, researchers in psychology consider human-centricity as an issue of job design where AI applications support operators' authority and wellbeing (e.g., De Cremer and Kasparov, 2021), which means that they combine organizational and individual characteristics. Researchers in engineering and manufacturing most likely address AI-based assistance to compensate for individual weaknesses in the production flow (Mehta et al., 2022) and thus relate technological and organizational characteristics to the individual, but with another concept of man than prevalent in psychology (Wilkens et al., 2021a). The number of coexisting definitions emphasizing different criteria can easily be interpreted as contradictory or controversial. We ask whether there is a system that allows us to relate different criteria to each other from a higher order. Reflecting on human-centricity requires a consideration of the perspectives on human-AI interaction (Anthony et al., 2023), the context characteristics of where AI is in use (Widder and Nafus, 2023), the individual demands of employees who are confronted with technology, and the responsibilities of stakeholders who are in charge of it (Polak et al., 2022). This is why we apply configurational theory (Mintzberg, 1993, 2023) to the meaning of the human-centricity of AI at work.

Basically, AI is a term for software applications dedicated to detecting patterns based on neural networks and various machine learning (ML) algorithms nowadays, aiming at copying human intelligence on a computational basis but without any parallel to human intelligence in terms of the underlying learning process (Wilkens, 2020; Russell and Norvig, 2021). The characteristics of AI evolve with the different waves of technology development (Launchbury, 2017; Xu, 2019), and definitions change accordingly. AI applications from the second wave of AI development can be described as pre-trained and fine-tuned machines having "the ability to reason and perform cognitive functions such as problem-solving, object and word recognition, and decision-making" (Hashimoto et al., 2018, p. 70). In the current third wave, scholars emphasize artificial general intelligence in terms of "intelligent agents that will match human capabilities for understanding and learning any intellectual task that a human being can" (Fischer, 2022, p. 1). Conversational Large Language Models give an example in this direction, and the high-speed dissemination of the non-licensed version of ChatGPT III shows that generative AI is not necessarily officially implemented in a work context but is prevalent due to high individual user acceptance, leading to continuous application in operational tasks. This challenges all fields of the private and public sectors and fosters the need to specify and reflect on the criteria of human-centricity against the background of technology development on the one hand and the characteristics of the use fields on the other. Current state-of-the-art research argues that there is a need for a contextualized understanding of AI at work and corresponding research methods (Anthony et al., 2023; Widder and Nafus, 2023). We transfer this consideration to the reflection on the human-centricity of AI, as the technology only belongs to work contexts while being promoted by a group of incumbents.

The research community in organization studies is well known for context-related distinctions, avoiding one-best-way or one-fits-all thinking. Scholars rather search for typologies under which conditions and characteristics matter most and thus lead to contextualized understandings of challenges and related performative practices (Miller, 1986; Mintzberg, 1993, 2023; Greckhamer et al., 2018). This consideration has already been applied to the first reflections on human-centered AI in work contexts (Wilkens et al., 2021b), but definitions of human-centricity often claim to be universal or at least disregard the contextual background they have been stated for. Our argument is that different definitions and criteria of human-centricity result from different research communities or peer groups with different use fields, functions, or responsibilities explicitly or implicitly in mind. This includes considerations like who is in charge of promoting a criterion in concrete developments and operations.

A configurational approach is proposed to be helpful in understanding from a higher order when a criterion of human-centricity is highlighted for generating solutions and when it can be subordinated or neglected in the face of specific context-based responsibility. Our aim of analysis is to identify typical configurations of human-centered AI in the organization and to specify and distinguish the meaning and relevance of human-centricity against the background of who is in charge of a specific work context. A deep understanding of context requires ethnographic research (Anthony et al., 2023; Widder and Nafus, 2023) but can be systematically prepared by a cross-disciplinary literature review, giving attention to contexts and determining which community emphasizes which criteria and why. This contributes to a common ground in theory development on human-centered AI as it enables systematizing various findings from the many research communities elaborating on this topic. It also provides practitioners with guidance in deciding which criteria matter most for which purpose and peer group and allows them to estimate when to focus on selected criteria and when to broaden their perspective while taking alternative views.

A reflection on human-centricity in connection with AI and work is a sociotechnical system perspective by its origin, as the three entities

of technology, human agency, and organization with their institutional properties are interrelated (Orlikowski, 1992; Strohm and Ulich, 1998). How a sociotechnical system perspective can be combined with a configurational approach will be outlined in the next section. In the third section, we explain the research method of a systematic literature review, including search strategy and data evaluation. Based on this, we outline the research findings first by an analytical distinction of eight criteria of human-centricity and, in the second step, by contextualizing and synthesizing them to seven configurations of actor-structure engagements. The concluding discussion and outlook feeds the results back to norming initiatives and emphasizes further empirical validation in future research.

2. Configurational perspective on human-centered AI in sociotechnical systems

Configurational theory is an approach among scholars in organizational studies that focuses on the distinction of typologies. Typologies are based on "conceptually distinct [organizational] characteristics that commonly occur together" (Meyer et al., 1993, p. 1175; see also Fiss et al., 2013). The analysis is related to equifinality by explaining episodic outcomes instead of separating between independent and dependent variables, which is nowadays also described as causal complexity by scholars promoting a neo-configurational approach (Misangyi et al., 2017). This is how and why configurational thinking is distinguished from contingency theory, which is drilled to find a context-related best fit between organizational practices and external demands (Meyer et al., 1993). Configurational theory calls for alternative qualitative research methods and initiates its own movement in data analysis (Fiss et al., 2013; Misangyi et al., 2017).

Mintzberg (1979, 1993, 2023) is one of the most well-known researchers in configurational theory, with a distinction between structurational configurations originally known as structure in fives (Mintzberg, 1979, 1993) and recently readjusted while giving more attention to stakeholders and agency in addition to structural characteristics. Mintzberg (2023) outlines seven configurations deduced from the impact of five actor groups in terms of operators, middle managers, C-level managers, support staff and analysts, experts for standardizing the technostructure, as well as organizational culture, and external stakeholders such as communities, governments, or unions.

The core idea is that organizations can activate different mechanisms of coordination, communication, standardization, decentralization, decision-making, and strategizing to gain outcomes and that there is no one best way to do it. The diagnosis and understanding of the organizational mechanisms of being performative are crucial for activating them. From a research point of view, it is interesting to note that organizations can, however, be clustered and distinguished by configurations that represent ideal types of success while gaining a specific organizational shape (Mintzberg, 1979, 2023).

The configurational theory was originally focused on the analysis and description of organizational characteristics but was also supposed to serve as a framework for the analysis of the individual and group level, respectively, a "sociotechnical systems approach to work group design" (Meyer et al., 1993, p. 1186; see also Suchman, 2012). This is exactly how Orlikowski (1992) explained sociotechnical systems with three interrelated entities: technology, human actors, and the organizational institutional context. From this perspective, technology is not a context-free object but is interpreted and enacted by human agents under organizational characteristics, which also leads to different meanings of technology when applied to and enacted in different settings (Orlikowski, 2000, 2007). The inseparability between social and technological entities was later described as entanglement and sociomateriality (Orlikowski and Scott, 2008; Leonardi, 2013).

However, configurational theory and methods are not very common in sociotechnical system analysis and can only be loosely applied by a few scholars (e.g., Pava, 1986; Badham, 1995). A reason might be that the approach gained great attention in organization studies but is often counterintuitive to the research methods applied in engineering and psychology, both disciplines with a strong emphasis on causality and linear thinking, which are adjoining disciplines elaborating on sociotechnical system thinking but with distinct research traditions and methods in use (Herzog et al., 2022). It is interesting to note that the detection of patterns is a mutual interest between ML approaches and organizational configurational theory but that the system-dynamic-based acyclic thinking of configurational theory is untypical of how ML methods currently work.

The reason we suggest elaborating on a configurational approach is that there is no single or prior group in charge of a human-centered AI application in work settings; instead, many disciplines and stakeholders involved from different levels of hierarchy and professions from inside and outside the organization contribute to the same topic. Consequently, there is a high plausibility that different approaches and stakeholders contribute to human-centricity and that there is no one best way or mastermind orchestration but different ways of enacting selected criteria dedicated to the human-centricity of AI at work. This is why we aim to explore these configurations and reflect them as a starting point to enhance the human-centricity of AI in organizations with respect to their contributions and limitations.

3. Literature review on the human-centricity of AI at work

3.1. Search strategy and data evaluation

To identify the most typical configurations in current academic writings and underlying fields of AI application, it is necessary to include a wide range of publications in the search strategy and to analyze the research contributions as systematically as possible. Since research on human-centered AI or work with AI is not limited to the management field but also includes disciplines such as work science, psychology, medicine, computer and information science, or even philosophy and sociology, we conduct a cross-disciplinary literature review with a systematic search strategy (Snyder, 2019). As a starting point, we use the 79 articles already identified from the review by Wilkens et al. (2021a), leading to the distinction of five criteria regarding trustworthiness and explainability, compensating individual deficits, protecting health, enhancing individual potential, and specifying responsibilities. Aligned with the guidelines of Page et al. (2021; see also Figure 1), we then systematically searched

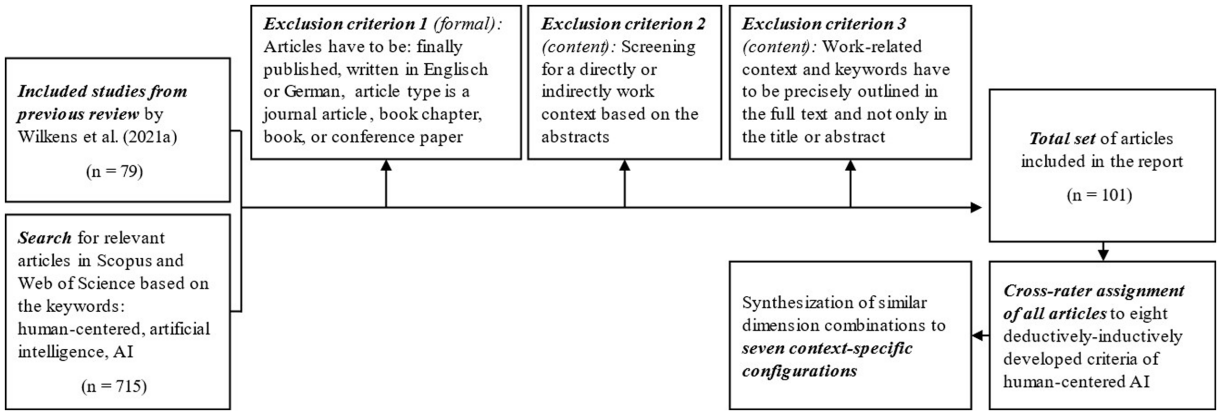

FIGURE 1
Flow diagram of the search strategy process, according to Page et al. (2021).

the Scopus and Web of Science databases for the keywords "human-centered" or "human" and "artificial intelligence" or "AI," as well as various synonyms, spellings, and their German translations. To consider the different publication strategies of the targeted disciplines, we included books, book chapters, journal articles, and conference papers and did not focus on discipline-specific journal ratings. By using boolean operators, we were able to identify a total of 715 additional articles. In the set of articles, we included all English and German language results but excluded articles in other languages that only had an English abstract or those that have not yet been published. In the second step, we screened all articles based on their abstracts and checked whether they contributed directly or indirectly to work to exclude those contributions with a pure focus on human-centered technology but without even an indirect reference to work. We also excluded papers with a pure interest in humanoid robots but without any interest in human-centered work. A human-technical focus facing technical design differed from a sociotechnical perspective and was therefore eliminated for the purpose of our analysis. However, the indirect reflection of work seemed to be of high relevance, which means that we did not exclude contributions when it became obvious that authors consider the technology relevant for future work settings or if they describe the work process of software development itself even though they do not name it work. In the third step, we delved deeper and analyzed the articles based on their full texts. We excluded all articles that only mentioned the relevant keywords in the title or abstract but did not discuss them in detail in the text. This search strategy resulted in a total set of 101 articles, of which 70 followed a theoretical-conceptual approach and 31 an empirical approach. Most of the authors of the articles were from the fields of computer science and engineering. However, due to the interdisciplinary scope of the articles, they were complemented by co-authors from the fields of management studies, psychology, ergonomics, and social science, as well as healthcare and education, to mention the most common backgrounds of co-authors.

Since we do not aim to quantify the literature but are interested in the underlying structure of its content, we followed a content analysis approach while analyzing the literature (Kraus et al., 2022). This involved reading the articles in their entirety by the authors and identifying dimensions of human-centered AI at work or human-centered work with AI. Therefore, the overall data evaluation process was twofold. The first step was analytical and aimed at the specification of dimensions and criteria indicating different meanings of human-centricity while working with AI. Here, we followed a deductive-inductive approach and used the five categories explored by Wilkens et al. (2021a) as deductive starting points and complemented and redefined them in several stages with cross-rater validation among all three authors by further inductively explored categories. Distinctions between categories are made when there are different meanings reflecting the underlying aim and intent of a human-centered approach. Homogeneity in intent and debate leads to a single category. Separable debates lead to the proposition of a further category (see Table 1). As a first result, we specified eight criteria for human-centered AI at work or working with AI.

The second step of analysis reflected the analytically separated criteria and synthesized them into seven configurations. The *synthesis* results from (1) the *coincidence of criteria* related to a *dominant criterion* while reflecting (2) the *actor groups in charge* of the application of the set of criteria. Publications were systematized and finally assigned to a configuration against this background. To give an illustration with selected examples for the treatment of dominant and supporting criteria: Weekes and Eskridge (2022a) emphasized technological characteristics for fostering *explainability* but went further in a second publication (Weekes and Eskridge, 2022b) on "Cognitive Enhancement of Knowledge Workers," in which they reflect *human agency and augmentation*. This is why the same authors can be represented in different configurations by different publications—in this case, in the *engagement for data and technostructure* with the first paper and the *engagement for proficiency* with the second paper. In their second study, Weekes and Eskridge (2022b) also referred to individual health and trustworthiness as subordinate criteria to the dominant one. Romero et al. (2016) overlapped with them in the overall set of criteria, but employees' *physical and mental health* were in the foreground, while the optimization of operational processes and human agency are considered supporting criteria. Therefore, Romero et al. (2016) were assigned to the *engagement for employees' wellbeing*. Even though authors overlap on two criteria, the focus of the article and the dominant perspective can differ (see Table 1).

The synthesis also includes the stakeholders in charge of a criterion and, respectively, the surrounding criteria. People in charge

TABLE 1 Assignment of the articles from the literature review to the human-centered AI criteria and their condensation into configurations.

	Challenges in human-centered technology development		*Challenges in human-centered employee development*			*Challenges in human-centered organizational development*			**Responsibility/ People in charge**	**Configuration**
	Trustworthiness, Privacy & Ethics	**Explainability**	**Job loss prevention**	**Physical & Mental Health**	**Human agency & Augmentation**	**Compensation of weaknesses in the system**	**Knowledge utilization from the user domain**	**Accountability & Safety Culture**		
Frissen (2022); Havrda and Rakova (2020); Jiang et al. (2021); Łapińska et al. (2021); Nakao (2022); Tzachor (2022); Zhang (2020)	x								AI developers	*None* *(n=7 articles)*
Adadi and Berrada (2018); Ehsan and Riedl (2020); Garcia-Magarino et al. (2019); Gillies et al. (2016); Gunning et al. (2019); Hayes and Moniz (2021); Heier (2020); Heier (2021); Hepenstal et al. (2021a); How et al. (2020b); Kahng (2019); Luštrek et al. (2021); Organ et al. (2021); Plass et al. (2022); Ploug and Holm (2020); Wang et al. (2021); Weekes and Eskridge (2022a); Xie et al. (2019); Zhu et al. (2018);		**x**							AI developers	**(1) Engagement for data and technostructure** *(n=23 articles)*
Bond et al. (2019); Krzywdzinski (2023); Rožanec (2021); Springer (2019)	x	**x**								
Adnan et al. (2020); Beede et al. (2020); Wilkens et al. (2019)					x	**x**			Front-line engineers (for shop-floor operators and processes)	**(2) Engagement for operational process optimization** *(n=16 articles)*
Foresti et al. (2020); Hynynen (1992); Mehta et al. (2022); Schaal (2007); Wei et al. (2022)						**x**				
Davagdorj et al. (2021); Grischke et al. (2020); How et al. (2020); How et al. (2020a)		x				**x**				
Schmidtler et al. (2015)				x		**x**	x			
Kim et al. (2022)	x					**x**				
Morrow et al. (2023)						**x**	x			
Gu et al. (2021); Jarrahi (2018); Nisser and Malanowski (2019)			**x**				x		Work councils (for shop-floor operators)	**(3) Engagement for operators' employment** *(n=10 articles)*
Guszcza et al. (2017), Kaiser and Malanowski (2019); Nahavandi (2019)			**x**		x					
Holzinger et al. (2022)		x	**x**				x			
Wilson and Daugherty (2018)		x	**x**							
Flandrin et al. (2021)			**x**		x	x				
Freitag et al. (2016)			**x**			x				
Romero et al. (2016)				**x**	x	x			HR, Ergonomists (for shop-floor operators)	**(4) Engagement for employees' well-being** *(n=11 articles)*
Riener et al. (2006); Taryudi et al. (2022)				**x**		x				
Del Guidice et al. (2023); Parker and Grote (2022); Kluge et al. (2021)				**x**	x			x		
Elkmann (2013); Hinds et al. (2004); Schönböck et al. (2022); Seabra et al. (2022)				**x**						
Riener et al. (2005)		x		**x**						
Weekes and Eskridge (2022b)	x			x	**x**				Professionals (experts)	**(5) Engagement for proficiency** *(n=19 articles)*
Akata et al. (2020); Lukowicz (2019)	x	x			**x**		x			
Dewey and Wilkens (2019)	x				**x**		x	x		
Hepenstal et al. (2021b); Holstein et al. (2019); Soldatos and Kyriazis (2021)	x				**x**					
Maiden et al. (2022)		x			**x**					
Ahrens (2014); Battaglia et al. (2021); Jung et al. (2022); Kathuria and Kathuria (2020); Smith et al. (2018); Steels (2020); Wilkens (2020)					**x**					
Guszcza (2018)		x			**x**					
Gamkrelidze et al. (2022)	x	x			**x**					
Kaasinen et al. (2022)	x				**x**					
Häusler and Sträter (2020)					**x**		x			
Ho et al. (2020); Shaikh (2020); Shneiderman (2020a); Shneiderman (2020c)	x							**x**	C-level management	**(6) Engagement for accountability** *(n=9 articles)*
Albahri et al. (2023); Barredo Arrieta (2020); Riedl (2019)	x	x						**x**		
Liu et al. (2022)								**x**		
Shneiderman (2020b)					x			**x**		
Bettoni (1995); Cui and Dai (2008); Fox et al. (2020); Mhlanga (2022); Polak et al. (2022)							**x**		Change management agents	**(7) Engagement for interactive cross-domain design** *(n=6 articles)*
Xu (2019)	x				x		**x**			

The bold font highlights the label of the seven configurations.

are not always explicitly mentioned but sometimes remain implicit. To gain access to the implicit assumptions, Mintzberg's organizational actors' description (Mintzberg, 2023, p. 17) serves as a blueprint for specifying the addressed audience. To add an illustration for this challenge, Shneiderman (2020a,b,c) stresses *accountability and safety culture* as important issues in human-centered AI, but without naming responsible actors. However, from a contextualized organizational understanding, it is obvious that this is an overall C-level responsibility and that the *top management team* can be specified as the actor in charge.

TABLE 2 Human-centered AI criteria and actor-structure engagements.

Challenges in human-centered technology development		*Challenges in human-centered employee development*			*Challenges in human-centered organizational development*			*Contextualized configuration*	*Responsibility/ Persons in charge*
Trustworthiness Privacy & Ethics	**Explainability**	**Physical & Mental Health**	**Human agency & Augmentation**	**Job loss prevention**	**Compensation of weaknesses in the system**	**Knowledge utilization from the user domain**	**Accountability & Safety Culture**		
15%	**85%**							(1) Engagement for data and technostructure	AI developers
4%	15%	4%	11%		**59%**	7%		(2) Engagement for operational process optimization	Front line engineers (for shop-floor operators and processes)
	9%		18%	**46%**	9%	18%		(3) Engagement for operators' employment	Work councils (for shop-floor operators)
	5%	**53%**	16%		16%		10%	(4) Engagement for employees' well-being	HR, Ergonomists (for shop-floor operators)
23%	13%	2%	**50%**			10%	2%	(5) Engagement for proficiency	Professionals (experts)
33%	17%		5%				**45%**	(6) Engagement for accountability	C-level management
12%			12%			**76%**		(7) Engagement for interactive cross-domain design	Change management agents

Percentages indicate the distribution of the human-centered AI criteria per configuration. For example, 23 articles are assigned to configuration (1) Engagement for data and technostructure (see Table 1). Of these, 19 refer exclusively to the criterion of Explainability and 4 to both Explainability and Trustworthiness, Privacy and Ethics. This total of 27 references results in a weighting of 85% for Explainability and 15% for Trustworthiness, Privacy, and Ethics in the configuration.
The shaded numbers highlight higher values. The highest values are shaded in dark and highlighted in bold.

Once the (1) *coincidence of criteria* and the (2) *actor groups in charge* are identified, it becomes apparent that the eight criteria of human-centered AI at work result in seven configurations. Considering the distribution of the criteria according to the frequency of their occurrence per configuration, the relative weighting reveals that each of the seven configurations is based on one dominant criterion, most likely surrounded by one or two other criteria, which reinforces the synthesis into seven configurations (see Table 2). Adding total numbers to the configurations, we observed that there were 23 reviewed publications with a core emphasis on the first configuration, the engagement for data and technostructure, 16 with an emphasis on the second engagement for operational process optimization, 10 with an emphasis on the third engagement for operators' employment, 11 with an emphasis on the fourth engagement for employees' wellbeing, 19 with an emphasis on the fifth engagement for proficiency, 9 with an emphasis on the sixth engagement for accountability, and 6 with an emphasis on the seventh engagement for interactive cross-domain design. A smaller number of publications related to a configuration does not indicate a lower relevance but only that there is currently less emphasis on the criterion or that the overall research community elaborating on a specific configuration is smaller. The differences in the distribution can rather be interpreted as a sign that relevant criteria, e.g., facing challenges in organizational development, can easily be overseen if the group of scholars representing them stands behind the dominant discourse with another emphasis, e.g., facing challenges in technology development.

3.2. Criteria of human-centered AI and how they lead to configurations

We identified eight criteria of human-centricity; two of them were discussed as challenges of human-centered technology development, three of them as challenges of human-centered employee development, and three of them as challenges of human-centered organizational development (see Table 1). A broader group of scholars asks how reliable and supportive AI-based technology is for individual decision-making and operations. They face the *challenges of human-centered technology development* with two criteria that are of key concern. The criterion of *trustworthiness, privacy, and ethics* means that the data structure is unbiased and that there is no ethical concern with respect to collecting and/or using the data. The goal is for AI to operate free from discrimination and provide reliable and ethical outcomes. The criterion of *explainability* means that the technology provides transparency about the data in use, how they are interpreted, and what error probability remains when using AI for decision support. The aim here is to enhance technology acceptance while giving helpful information to the user. Even though both criteria relate to the same challenges of the data structure, which is why they were comprised by Wilkens et al. (2021a), the underlying aim and intent differ in such a way that we propose to treat them separately.

Another group of scholars faces the *challenges of human-centered employee development*. The coding process explored three criteria. The first criterion results from an overall debate primarily addressed in social science. It is the *prevention of job loss*. Empirical findings show that new technologies, as well as digitalization and AI, lead to an increase in jobs at the level of economies, and a specific group of jobs, e.g., standardized tasks in manufacturing, logistics, or administration, can be reduced (Petropoulos, 2018; Arntz et al., 2020). As a single employee or group of employees might suffer these effects, the criterion can matter at the company level, which leads to the discourse of preventing employees from negative consequences due to new technologies. With the criterion of *physical and mental health,* scholars give emphasis to the protection of employees while aiming at preventing them from negative influences such as heavy loads, chemical substances, stressful interactions, etc., which they have to cope with while performing operational tasks. This is a group of scholars with a background in ergonomics and a stable category that already occurred in the review from Wilkens et al. (2021a). The criterion of *human agency and augmentation* is a further stable outcome of the coding process. The category is taken into

consideration across certain disciplines. The meaning is to design and use technology in such a manner that employees are in control of the technology (Legaspi et al., 2019) while performing tasks in direct interaction with AI and experiencing empowerment and further professionalization through the human-AI interaction.

A third overall dimension is related to the *challenges of human-centered organizational development*. The meaning of human-centricity is to reflect human needs and potentials, as well as weaknesses and negligence, to keep systems and interactions going and make them safe and reliable. One criterion is the *compensation of weaknesses and system optimization*. This explores a rather deficit-oriented perspective on the human being because of fatigue, unstable concentration, or limits in making distinctions on the basis of human sensors. AI is considered an approach to compensate for these weaknesses (Wilkens et al., 2021a). However, this is not for drawing a rather negative picture of the human being but to keep the system working and optimize processes where there would otherwise be negative system outcomes. The aim of this human-centered approach is high precision, failure reduction, high speed, and high efficiency. The criterion of *integration of user domain knowledge* gives attention to the connection between the domain of software development and the user domain. More traditionally, this is user-centered design and tool development, an approach that has been advocated for almost 30 years (see Fischer and Nakakoji, 1992). In current further development, it is not primarily the end-user need but the integration of user domain knowledge in the software development process to make the technology better and more reliable on the system level due to feedback loops between these domains and the expertise resulting from user domain knowledge. The clue is higher proficiency in technology development through job design principles across domains. Finally, there is the criterion of *accountability and safety culture* based on the meaning of human-centricity: a long-term benefit from AI requires reliable systems and organizational routines that guarantee this reliability. The goal is to provide and implement clear process descriptions and checklists that foster high levels of responsibility at the system level.

These eight criteria related to three dimensions can comprise seven contextualized configurations of an actor-structure engagement, specifying who is in charge of fostering what criteria, the (acceptable) limitations of the approach, and the need to elaborate on a broader view of the system level. While all seven configurations are each based on a dominant criterion, one criterion represents an exception. Trustworthiness, privacy, and ethics support almost all configurations and can thus be classified as a necessary overall condition (see Table 2; Wilkens et al., 2021b).

Note: Percentages indicate the distribution of the human-centered AI criteria per configuration. The weighting is based on the absolute number of articles assigned to the dimensions.

The configuration (1) *engagement for data and technostructure* identified from 23 publications under leading authorship from computer science is based on the criterion *explainability* of AI and is often brought by AI developers in charge of technical applications from outside the user domain to the specific workplace. This criterion is supported by trustworthiness, privacy, and ethics. The impact from outside the organization includes a wide range of industries, from manufacturing, business, healthcare, and education to the public sector. The quality of the technology itself is an issue of human-centricity, but without reflecting other criteria with respect to the employee or organizational development of the absorbing organizations. This means that high-end technology affects the standards and technostructure of other organizations without considering the consequences. However, those who develop technology have a guideline for keeping the developed tool's quality as high as possible.

The second configuration detected from 16 publications is the (2) *engagement for operational process optimization*. Those who are in charge face the challenges of organizational development with respect to operators' workflows. The primary criterion is the *compensation of weaknesses* for high system outcomes in terms of accuracy, quality, and efficiency. Authors in engineering are prevalent in this class. A combination of employee development-related criteria occurs in some writings, but the contextualized approach is dedicated to process design. The responsibility is especially taken by line management engineers who follow design principles for optimizing system outcomes while compensating for human weaknesses with the help of sophisticated technology.

The third configuration is (3) *engagement for operators' employment* with a key criterion of *preventing employees*, especially front-line shop-floor operators, *from job loss,* which could be explored in 10 publications from interdisciplinary author groups. This approach to human-centricity is often discussed as the back side of the medal when the technostructure or the optimization of operational processes—both configurations were just outlined—are considered in an isolated manner. This perspective gives prior emphasis to employee development and is also surrounded by further criteria related to technology or organizational development. Those who are in charge, e.g., work councils from inside the organization or unions from outside, aim at keeping employment within a company high—often not just as a means but also as an end. Those who feel responsible for keeping employment high within the company have a starting point for their inquiry and also an approach to further criteria fostering operators' employment.

The fourth configuration prevalent in 11 publications is the (4) *engagement for employees' wellbeing*, emphasizing *physical and mental health*, especially of operators. Co-authors represent this expertise. Their focus is enriched by further criteria related to employee or organizational development. Technology is often not specified in this configuration but is prevalent as an initial point to reflect on human-centricity. Another crucial point is that the whole job profile—and not just a single task—is reflected against the background of AI applications. The groups proposed to be in charge of this configuration are HR staff members or ergonomists.

With the fifth configuration, (5) *engagement for proficiency*, deduced from 19 publications with authors from a wide range of disciplines, the focus shifts from operators, often considered shop-floor operators, to different individual experts within the organization who are responsible for decision-making and solutions with critical impact, e.g., in medical diagnosis, surgeries, or business development. These experts are often at the medium or top level within the organization. The key criterion is *human agency and augmentation,* most likely supported by the criteria of trustworthiness and explainability of AI. The issue is hybrid intelligence for specific tasks and decisions, not necessarily whole job profiles. The addressed experts are often not organized by others or confronted with new technology but decide its application themselves. This is why they can

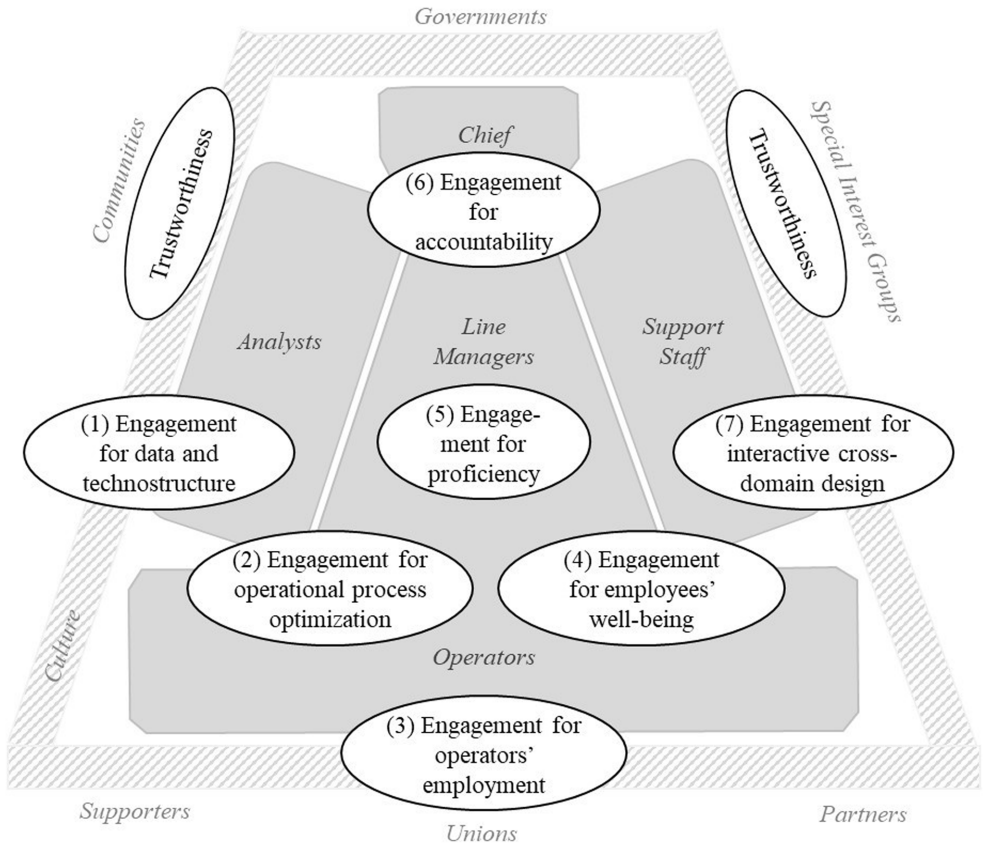

FIGURE 2
Actor-structure engagements of human-centered AI in organizational contexts. Adaptation of Mintzberg's organizational actors (Mintzberg, 2023, p. 17, gray structure in the background; use of figure authorized by Mintzberg via email) by the actor-structure engagements of human-centered AI explored in the literature review (white cells).

focus on the quality of the technology and the outcome for their individual profession, often at the middle level of an organization.

The sixth configuration, (6) *engagement for accountability*, with an underlying number of nine publications from different disciplines, further shifts the focus to the C-level managers in charge of decisions affecting the overall organizational development. It is the *accountability and safety culture*, especially at critical interfaces within and across organizations, that is the key criterion for this configuration of human-centricity. The criterion is often enriched by the trustworthiness of the AI application. This underlines that the top management team pursues other criteria of human-centricity than, e.g., the work councils or HR managers.

The final configuration was detected in six publications situated in different disciplines: (7) *Engagement for interactive cross-domain design* faces another challenge of organizational development: *knowledge utilization from the user domain* in the process of AI development. This perspective currently gains great attention in co-creation and co-design research (Russo-Spena et al., 2019; Li et al., 2021; Suh et al., 2021). In the search field of human-centered AI, the perspective is rather new and currently leads to a bi-directional exchange of knowledge to reach high reliability and safety for AI applications. This configuration is of key concern for work processes in software development companies and user domain firms. It is especially organizational development or change management experts who take responsibility for this perspective and criterion. This configuration builds bridges to the first configuration and aims at AI applications that are adaptable to a firm's standards and technostructure and thus also avoid negative side effects, as especially anticipated in the second and third configurations of operational process optimization and operators' employment.

A configuration is related to the fields of responsibility of organizational internal or external stakeholders who are in charge of human-centered outcomes in a sub-field of an overall process or design. This is why the identified configurations can be specified and aligned to Mintzberg's (2023) actor-structure constellations (see Figure 2). The search for configurations revealed that no mastermind covers all criteria when contextualizing the human-centricity of AI at work but that each criterion needs to be advocated by responsible stakeholders. This leads to distinct approaches across hierarchy and expertise within organizations and makes it challenging to fulfill the overall mission of human-centered AI in the workplace.

However, the actor-structure engagement for selected criteria is a feasible approach for those with responsibilities and promotes the development as long as the stakeholders acknowledge additional perspectives and contributions from other domains or positions. To get to more integrative solutions, a first step could be to align two or three configurations with each other, e.g., the engagement for data and technostructure with the engagement for proficiency. This is especially

helpful when they complement each other, e.g., the engagement for interactive cross-domain design allows to cope with the limitations that go hand in hand with the engagement for data and technostructure and to foster employees' wellbeing.

4. Discussion, limitations, and outlook on future research

The systematic literature review across certain disciplines explores criteria of human-centricity while integrating AI in the work context. We could identify a variety of criteria that either face challenges of human-centered technology development, human-centered employee development, or human-centered organizational development. With this distinction, we could further develop already existing classifications (Wilkens et al., 2021a) and substantiate that the reflection on AI at work goes beyond issues of human-technology interaction but also includes organizational processes, structures, and policies. A further advancement is the synthesis of the eight analytically distinguishable criteria into seven context-related configurations, specifying the actor-structure engagement behind these criteria. Depending on the organizational sub-unit and the typical stakeholders involved in that unit, one criterion takes precedence and is supported by other criteria, while other criteria tend to be neglected. Considering the identified engagements for human-centricity against Mintzberg's (2023) model of organizational configurations, it becomes obvious that all structural parts and related actors—operators, line managers, C-level managers, analysts, and support staff—are involved and in charge. The identified eight criteria of human-centricity and seven configurations of enacting and contextualizing them complement each other meaningfully and lead to a holistic overall approach. However, there is no actor-structure configuration, including all criteria, as a kind of mastermind approach.

Comparing the prevalent criteria of human-centricity as deduced from the academic discourse with the proposals for responsible AI declarations and regulations, it becomes obvious that outlines such as the EU AI Act (European Parliament, 2023) primarily face the two challenges of human-centered technology development. This is also the case for the industry norm ISO/IEC TR 24028:2020 (2020). Interestingly, the recently published proposal of the US Blueprint AI Bill of Rights goes beyond and considers the integration of user-domain knowledge in the AI development process and operators' wellbeing as crucial points in addition to technology development (The White House, 2022). The industry norm ISO 9241-210:2019 (2019) gives emphasis to physical and mental health, especially mental load while interacting with technology. Even though the norm does not address AI explicitly, it can serve as a guideline for standards as long as more specific AI-related norms for human-AI interaction are missing. However, it also becomes obvious that other challenges of human-centered employee development and human-centered organizational development, especially with respect to human agency and augmentation and related process descriptions in job design, are neglected in comparison to the more traditional outlines of human wellbeing. This will be a future task. There is a rising number of organizations such as Microsoft, SAP, Bosch, or Deutsche Telekom that have company guidelines or codex agreements (Deutsche Telekom, 2018; Robert Bosch, 2020; SAP SE, 2021; Microsoft Corporation, 2022). They tend to include challenges of technology, employee, and organizational development but, at the same time, tend to be more vague in what criteria are addressed. However, it is interesting to note that accountability and safety culture gain attention in these declarations at the company level. This underlines C-level responsibility in the overall firm strategy. To date, only a few companies have published these guidelines. Future research will have to compare in more detail which criteria elaborate on an industry norm or are even an issue of legal regulation that tends to remain in the background and what the implications are when criteria are weighted unequally.

The overall implication of the norming initiatives is that, from an organizational actor perspective, these standards are supposed to be integrated into organizations by stakeholders from the legal departments, almost belonging to the support staff. Consequently, this group of stakeholders might have a higher impact in the future. While AI developers in the scientific discourse are in charge of the criteria due to formalization and regulation, they will rather be represented by lawyers in the practical context. This group of stakeholders could not be identified in such a clear manner from the conducted literature review. A higher engagement of lawyers, which can be expected in the future, can further foster the emphasis on human-centricity on the one hand.

On the other hand, this bears the risk that other criteria of human-centricity outlined in this review with a stronger emphasis on employee development and organizational development, which are less standardized so far, tend to be neglected or that the responsibility for human-centricity is delegated to the legal departments in organizations and not located where the AI development takes place (see Widder and Nafus, 2023). At least, there is a risk of overemphasizing technology-related criteria in comparison to the broader view provided in this article. A coping strategy could be to consider the technology-related criteria of human-centered AI as a necessary condition and to add on sufficient conditions related to the specific use field as proposed in the maturity model by Wilkens et al. (2021b).

The criteria and configurations explored in the systematic literature review need further empirical validation in the next step. This validation includes the analytical distinction of the named criteria and the context-specific consistency of the proposed configurations. Moreover, an empirical analysis should elaborate on further operationalizing the assumed related performative practices and outcomes. Another issue of empirical validation is to test whether configurations lead to a holistic perspective when integrating them or if there are shortcomings or differences due to power differences among the representing stakeholders, probably leading to crowding-out effects. The preferred approach for data evaluation is qualitative comparative analysis (QCA), as it is a mature concept especially developed for exploring configurations (Miller, 1986, 2017; Fiss et al., 2013; Misangyi et al., 2017).

The aim of the presented review was to elaborate on a common ground in human-centered AI at work, with an emphasis on the academic debate. The value and uniqueness of the approach lie in

the contextualization of criteria and the stakeholders in charge of them. This allows us to better understand how human-centricity belongs to the work context while being enacted by a group of stakeholders. This also explains the co-existence of different engagements for human-centricity and that this can even generate an advantage as long as the criteria complement and do not crowd out each other.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

UW: Data curation, Investigation, Methodology, Validation, Writing – original draft, Conceptualization, Project administration. DL: Data curation, Investigation, Methodology, Validation, Writing – original draft, Visualization. VL: Conceptualization, Data curation, Investigation, Project administration, Validation, Writing - review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The study was funded by Competence Center HUMAINE: Transfer-Hub of the Ruhr Metropolis for human-centered work with AI (human-centered AI network), Funding code: BMBF 02L19C200.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Adadi, A., and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access* 6, 52138–52160. doi: 10.1109/ACCESS.2018.2870052

Adnan, H. S., Matthews, S., Hackl, M., Das, P. P., Manaswini, M., Gadamsetti, S., et al. (2020). Human centered AI design for clinical monitoring and data management. *Eur. J. Pub. Health* 30:86. doi: 10.1093/eurpub/ckaa165.225

Ahrens, V. (2014). Industrie 4.0: Ein humanzentrierter Ansatz als Gegenentwurf zu technikzentrierten Konzepten. Working paper der Nordakademie Nr. 2014-05. Elmshorn.

Akata, Z., Balliet, D., Rijke, M., Dignum, F., Dignum, V., Eiben, G., et al. (2020). A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer* 53, 18–28. doi: 10.1109/MC.2020.2996587

Albahri, A. S., Duhaim, A. M., Fadhel, M. A., Alnoor, A., Baqer, N. S., Alzubaidi, L., et al. (2023). A systematic review of trustworthy and explainable artificial intelligence in healthcare: assessment of quality, bias risk, and data fusion. *Informat. Fusion* 96, 156–191. doi: 10.1016/j.inffus.2023.03.008

Anthony, C., Bechky, B. A., and Fayard, A. (2023). "Collaborating" with AI: taking a system view to explore the future of work. *Organ. Sci.* 34, 1672–1694. doi: 10.1287/orsc.2022.1651

Arntz, M., Gregory, T., and Zierahn, U. (2020). Digitalisierung und die Zukunft der Arbeit. *Wirtschaftsdienst* 100, 41–47. doi: 10.1007/s10273-020-2614-6

Badham, R. (1995). "Managing sociotechnical change: a configuration approach to technology implementation," in *The symbiosis of work and technology*, ed. J. Benders, HaanJ. de and D. Bennett London: Taylor & Francis Ltd., 77–94.

Barredo, A. A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A. B. A. A., et al. (2020). Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58, 82–115. doi: 10.1016/j.inffus.2019.12.012

Battaglia, E., Boehm, J., Zheng, Y., Jamieson, A. R., Gahan, J., and Fey, A. M. (2021). Rethinking autonomous surgery: focusing on enhancement over autonomy. *Eur. Urol. Focus* 7, 696–705. doi: 10.1016/j.euf.2021.06.009

Beede, E., Baylor, E., Hersch, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P., et al. (2020). "A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy" in *Proceedings of the 2020 CHI conference on human factors in computing systems* (New York, NY: Association fpr Computing Machinery), 1–12.

Bettoni, M. C. (1995). Kant and the software crisis: suggestions for the construction of human-centred software systems. *AI & Soc.* 9, 396–401. doi: 10.1007/BF01210590

Bond, R., Mulvenna, M. D., Wan, H., Finlay, D. D., Wong, A., Koene, A., et al. (2019). "Human centered artificial intelligence: weaving UX into algorithmic decision making," In *2019 16th international conference on human-computer interaction (RoCHI)* (Bucharest, RO), 2–9.

Cui, X., and Dai, R. (2008). A human-centred intelligent system framework: meta-synthetic engineering. *International Journal of Intelligent Information and Database Systems* 2, 82–105. doi: 10.1504/IJIIDS.2008.017246

Davagdorj, K., Bae, J. W., Pham, V. H., Theera-Umpon, N., and Ryu, K. H. (2021). Explainable artificial intelligence based framework for non-communicable diseases prediction. *IEEE Access* 9, 123672–123688. doi: 10.1109/access.2021.3110336

De Cremer, D., and Kasparov, G. (2021). AI should augment human intelligence, not replace it. *Harv. Bus. Rev.* 18

Del Giudice, M., Scuotto, V., Orlando, B., and Mustilli, M. (2023). Toward the human–centered approach. A revised model of individual acceptance of AI. *Hum. Resour. Manag. Rev.* 33:100856. doi: 10.1016/j.hrmr.2021.100856

Deutsche Telekom, AG (2018). *Digital ethics guidelines on AI*. Available at: https://www.telekom.com/resource/blob/544508/ca70d6697d35ba60fbcb29aeef4529e8/dl-181008-digitale-ethik-data.pdf

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

Ehsan, U., and Riedl, M. O. (2020). "Human-centered explainable AI: towards a reflective sociotechnical approach" in *2020 international conference on human-computer interaction (HCII)* (Copenhagen, DK: Springer, Cham), 449–466.

Elkmann, N. (2013). Sichere Mensch-Roboter-Kooperation: Normenlage, Forschungsfelder und neue Technologien. *Zeitschrift für Arbeitswissenschaft* 67, 143–149. doi: 10.1007/BF03374401

European Parliament (2023). *Artificial intelligence act, P9_TA(2023)0236*. Available at: https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_DE.html

Fischer, G. (2022). A research framework focused on 'AI and humans' instead of 'AI versus humans. *Interaction Design and Architecture(s) – IxD&A Journal.* doi: 10.55612/s-5002-000

Fischer, G., and Nakakoji, K. (1992). Beyond the macho approach of artificial intelligence: empower human designers – do not replace them. *Knowledge-Based Systems Journal, Special Issue on AI in Design* 5, 15–30. doi: 10.1016/0950-7051(92)90021-7

Fiss, P. C., Cambré, B., and Marx, A. (2013). "Configurational theory and methods in organizational research" in *Research in the sociology of organizations*. eds. P. C. Fiss, B. Cambré and A. Marx (Emerald Group Publishing Limited), 1–22.

Flandrin, P., Hellemans, C., Van der Linden, J., and Van de Leemput, C. (2021). "Smart technologies in hospitality: effects on activity, work design and employment. A case study about chatbot usage" in *2021 17th proceedings of the 17th "Ergonomie et Informatique Avancée" conference* (New York, NY: Association for Computing Machinery), 1–11. doi: 10.1145/3486812.3486838

Foresti, R., Rossi, S., Magnani, M., Bianco, C. G. L., and Delmonte, N. (2020). Smart society and artificial intelligence: big data scheduling and the global standard method applied to smart maintenance. *Engineering* 6, 835–846. doi: 10.1016/j.eng.2019.11.014

Fox, J., South, M., Khan, O., Kennedy, C., Ashby, P., and Bechtel, J. (2020). OpenClinical. net: artificial intelligence and knowledge engineering at the point of care. *BMJ Health Care Informatics* 27. doi: 10.1136/bmjhci-2020-100141

Freitag, M., Molzow-Voit, F., Quandt, M., and Spöttl, G. (2016). "Aktuelle Entwicklung der Robotik und ihre Implikationen für den Menschen" in *Robotik in der Logistik: Qualifizierung für Fachkräfte und Entscheider*. eds. F. Molzow-Voit, M. Quandt, M. Freitag and G. Spöttl (Wiesbaden, GE: Springer Fachmedien Wiesbaden), 9–20.

Frissen, V. (2022). "Working with big data and Al: toward balanced and responsible working practices" in *Digital innovation and the future of work*. eds. H. Schaffers, M. Vartiainen and J. Bus (New York, NY: River Publishers), 111–136. doi: 10.1201/9781003337928

Gamkrelidze, T., Zouinar, M., and Barcellini, F. (2022). "Artificial intelligence (AI) in the workplace: a study of stakeholders' views on benefits, issues and challenges of AI systems" in *Proceedings of the 21st congress of the international ergonomics association (IEA 2021)* (Cham: Springer International Publishing), 628–635.

Garcia-Magarino, I., Muttukrishnan, R., and Lloret, J. (2019). Human-Centric AI for trustworthy IoT systems with explainable multilayer perceptions. *IEEE Access* 7, 125562–125574. doi: 10.1109/ACCESS.2019.2937521

Gillies, M., Lee, B., d'Alessandro, N., Tilmanne, J., Kulesza, T., Caramiaux, B., et al. (2016). "Human-Centred machine learning" in *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems* (New York, NY: Association for Computing Machinery), 3558–3565.

Greckhamer, T., Furnari, S., Fiss, P. C., and Aguilera, R. V. (2018). Studying configurations with qualitative comparative analysis: best practices in strategy and organization research. *Strateg. Organ.* 16, 482–495. doi: 10.1177/1476127018786487

Grischke, J., Johannsmeier, L., Eich, L., Griga, L., and Haddadin, S. (2020). Dentronics: towards robotics and artificial intelligence in dentistry. *Dent. Mater.* 36, 765–778. doi: 10.1016/j.dental.2020.03.021

Gu, H., Huang, J., Hung, L., and Chen, X. A. (2021). Lessons learned from designing an AI-enabled diagnosis tool for pathologists. *Proc. ACM Hum. Comput. Interact.* 5, 1–25. doi: 10.1145/3449084

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., and Yang, G. Z. (2019). XAI—Explainable artificial intelligence. *Sci. Robot.* 4, 1–2. doi: 10.1126/scirobotics.aay7120

Guszcza, J. (2018). Smarter together: why artificial intelligence needs human-centered design. *Deloitte Rev.* 22, 36–45.

Guszcza, J., Harvey, L., and Evans-Greenwood, P. (2017). Cognitive collaboration: why humans and computers think better together. *Deloitte Rev.* 20, 7–30.

Hashimoto, D. A., Rosman, G., Rus, D., and Meireles, O. R. (2018). Artificial intelligence in surgery: promises and perils. *Ann. Surg.* 268, 70–76. doi: 10.1097/sla.0000000000002693

Häusler, R., and Sträter, O. (2020). "Arbeitswissenschaftliche Aspekte der Mensch-Roboter-Kollaboration" in *Mensch-Roboter-Kollaboration*. ed. H. J. Buxbaum (Wiesbaden, GE: Springer Fachmedien Wiesbaden), 35–54.

Havrda, M., and Rakova, B. (2020). "Enhanced wellbeing assessment as basis for the practical implementation of ethical and rights-based normative principles for AI" in *2020 IEEE international conference on systems, man, and cybernetics (SMC)* (Toronto, ON: IEEE), 2754–2761.

Hayes, B., and Moniz, M. (2021). "Trustworthy human-centered automation through explainable AI and high-fidelity simulation" in *2020 international conference on applied human factors and ergonomics (AHFE)* (Cham: Springer), 3–9.

Heier, J. (2021). "Design intelligence-taking further steps towards new methods and tools for designing in the age of AI" in *Artificial intelligence in HCI: Second international conference, AI-HCI international conference (HCII)* (Cham: Springer), 202–215.

Heier, J., Willmann, J., and Wendland, K. (2020). "Design intelligence-pitfalls and challenges when designing AI algorithms in B2B factory automation" in *Artificial intelligence in HCI: first international conference* (Cham: Springer), 288–297.

Hepenstal, S., Zhang, L., Kodagoda, N., and Wong, B. W. (2021b). "A granular computing approach to provide transparency of intelligent systems for criminal investigations" in *Interpretable artificial intelligence: a perspective of granular computing*. eds. W. Pedrycz and S.-M. Chen, vol. *937* (Cham: Springer)

Hepenstal, S., Zhang, L., and Wong, B. W. (2021a). "An analysis of expertise in intelligence analysis to support the design of human-centered artificial intelligence," in 2021 IEEE international conference on systems, man, and cybernetics (SMC) Melbourne, Australia (IEEE), 107–112

Herzog, M., Wilkens, U., Bülow, F., Hohagen, S., Langholf, V., Öztürk, E., et al. (2022). "Enhancing digital transformation in SMEs with a multi-stakeholder approach" in *Digitization of the work environment for sustainable production*. ed. P. Plapper (Berlin, GE: GITO Verlag), 17–35.

Hinds, P., Roberts, T., and Jones, H. (2004). Whose job is it anyway? A study of human-robot interaction in a collaborative task. *Hum. Comput. Interact.* 19, 151–181. doi: 10.1207/s15327051hci1901%262_7

Ho, C. W., Ali, J., and Caals, K. (2020). Ensuring trustworthy use of artificial intelligence and big data analytics in health insurance. *Bull. World Health Organ.* 98, 263–269. doi: 10.2471/blt.19.234732

Holstein, K., McLaren, B. M., and Aleven, V. (2019). "Designing for complementarity: teacher and student needs for orchestration support in AI-enhanced classrooms" in *Proceedings of the 20th international conference on artificial intelligence and education* (New York: Springer International Publishing), 157–171.

Holzinger, A., Saranti, A., Angerschmid, A., Retzlaff, C. O., Gronauer, A., Pejakovic, V., et al. (2022). Digital transformation in smart farm and forest operations needs human-centered AI: challenges and future directions. *Sensors* 22:3043. doi: 10.3390/s22083043

How, M.-L., and Chan, Y. J. (2020). Artificial intelligence-enabled predictive insights for ameliorating global malnutrition: a human-centric AI-thinking approach. *AI* 1, 68–91. doi: 10.3390/ai1010004

How, M.-L., Cheah, S.-M., Chan, Y. J., Khor, A. C., and Say, E. M. P. (2020a). Artificial intelligence-enhanced decision support for informing global sustainable development: a human-centric AI-thinking approach. *Information* 11, 1–24. doi: 10.3390/info11010039

How, M.-L., Cheah, S.-M., Khor, A. C., and Chan, Y. J. (2020b). Artificial intelligence-enhanced predictive insights for advancing financial inclusion: a human-centric AI-thinking approach. *BDCC* 4, 1–21. doi: 10.3390/bdcc4020008

Hynynen, J. (1992). Using artificial intelligence technologies in production management. *Comput. Ind.* 19, 21–35. doi: 10.1016/0166-3615(92)90004-7

ISO 9241-210:2019 (2019). Ergonomics of human-system interaction — Part 210: human-centred design for interactive systems.

ISO/IEC TR 24028:2020 (2020). Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence.

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: human-AI symbiosis in organizational decision making. *Bus. Horiz.* 61, 577–586. doi: 10.1016/j.bushor.2018.03.007

Jiang, J., Karran, A. J., Coursaris, C. K., Léger, P. M., and Beringer, J. (2021). "A situation awareness perspective on human-agent collaboration: tensions and opportunities" in *HCI international 2021-late breaking papers: multimodality, eXtended reality, and artificial intelligence: 23rd HCI international conference (HCII)*, vol. *39* (New York: Springer International Publishing), 1789–1806.

Jung, M., Werens, S., and von Garrel, J. (2022). Vertrauen und Akzeptanz bei KI-basierten, industriellen Arbeitssystemen. *Zeitschrift für wirtschaftlichen Fabrikbetrieb* 117, 781–783. doi: 10.1515/zwf-2022-1134

Kaasinen, E., Anttila, A. H., Heikkilä, P., Laarni, J., Koskinen, H., and Väätänen, A. (2022). Smooth and resilient human–machine teamwork as an industry 5.0 design challenge. *Sustainability* 14:2773.b. doi: 10.3390/su14052773

Kahng, M. B. (2019). *Human-centered AI through scalable visual data analytics*. [dissertation] [Georgia]: Institute of Technology

Kaiser, O. S., and Malanowski, N. (2019). Smart Data und Künstliche Intelligenz: Technologie, Arbeit, Akzeptanz. *Working Paper Forschungsförderung* 136. Available at: http://hdl.handle.net/10419/216056

Kathuria, R., and Kathuria, V. (2020). "The use of human-centered AI to augment the health of older adults" in *HCI international 2020–late breaking posters: 22nd international conference, HCII 2020* (Cham: Springer International Publishing), 469–477.

Kim, J.-W., Choi, Y.-L., Jeong, S.-H., and Han, J. (2022). A care robot with ethical sensing system for older adults at home. *Sensors* 22:7515. doi: 10.3390/s22197515

Kluge, A., Ontrup, G., Langholf, V., and Wilkens, U. (2021). Mensch-KI-Teaming: Mensch und Künstliche Intelligenz in der Arbeitswelt von morgen. *Zeitschrift für wirtschaftlichen Fabrikbetrieb* 116, 728–734. doi: 10.1515/zwf-2021-0112

Kraus, S., Breier, M., Lim, W. M., Dabić, M., Kumar, S., Kanbach, D. K., et al. (2022). Literature reviews as independent studies: guidelines for academic practice. *Rev. Manag. Sci.* 16, 2577–2595. doi: 10.1007/s11846-022-00588-8

Krzywdzinski, M., Gerst, D., and Butollo, F. (2023). Promoting human-centred AI in the workplace. Trade unions and their strategies for regulating the use of AI in Germany. *Transfer* 29, 53–70. doi: 10.1177/10242589221142273

Łapińska, J., Escher, I., Górka, J., Sudolska, A., and Brzustewicz, P. (2021). Employees' trust in artificial intelligence in companies: the case of energy and chemical industries in Poland. *Energies* 14:1942. doi: 10.3390/en14071942

Launchbury, J. (2017). *A DARPA perspective on artificial intelligence*. DARPA talk, February 15, 2017.

Legaspi, R., He, Z., and Toyoizumi, T. (2019). Synthetic agency: sense of agency in artificial intelligence. *Curr. Opin. Behav. Sci.* 29, 84–90. doi: 10.1016/j.cobeha.2019.04.004

Leonardi, P. M. (2013). When does technology use enable network change in organizations? A comparative study of feature use and shared affordances. *MIS Q.* 37, 749–775. doi: 10.25300/MISQ/2013/37.3.04

Li, S., Peng, G., Xing, F., Zhang, J., and Qian, Z. (2021). Value Co-creation in Industrial AI: the interactive role of B2B supplier, customer and technology provider. *Ind. Mark. Manag.* 98, 105–114. doi: 10.1016/j.indmarman.2021.07.015

Liu, C., Tian, W., and Kan, C. (2022). When AI meets additive manufacturing: challenges and emerging opportunities for human-centered products development. *J. Manuf. Syst.* 64, 648–656. doi: 10.1016/j.jmsy.2022.04.010

Lukowicz, P. (2019). The challenge of human centric. *Digitale Welt* 3, 9–10. doi: 10.1007/s42354-019-0200-0

Luštrek, M., Bohanec, M., Barca, C. C., Ciancarelli, M. G. T., Clays, E., Dawodu, A. A., et al. (2021). A personal health system for self-management of congestive heart failure (HeartMan): development, technical evaluation, and proof-of-concept randomized controlled trial. *JMIR Med. Inform.* 9:e24501. doi: 10.2196/24501

Maiden, N., Lockerbie, J., Zachos, K., Wolf, A., and Brown, A. (2022). Designing new digital tools to augment human creative thinking at work: an application in elite sports coaching. *Expert. Syst.* 40. doi: 10.1111/exsy.13194

Mehta, R., Moats, J., Karthikeyan, R., Gabbard, J., Srinivasan, D., Du, E., et al. (2022). Human-centered intelligent training for emergency responders. *AI Mag.* 43, 83–92. doi: 10.1002/aaai.12041

Meyer, A. D., Tsui, A. S., and Hinings, C. R. (1993). Configurational approaches to organizational analysis. *Acad. Manag. J.* 36, 1175–1195. doi: 10.5465/256809

Mhlanga, D. (2022). Human-centered artificial intelligence: the superlative approach to achieve sustainable development goals in the fourth industrial revolution. *Sustainability* 14:7804. doi: 10.3390/su14137804

Microsoft Corporation (2022). *Microsoft responsible AI standard, v2*. Available at: https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE5cmFl

Miller, D. (1986). Configurations of strategy and structure: towards a synthesis. *Strateg. Manag. J.* 7, 233–249. doi: 10.1002/smj.4250070305

Miller, D. (2017). Challenging trends in configuration research: where are the configurations? *Strateg. Organ.* 16, 453–469. doi: 10.1177/1476127017729315

Mintzberg, H. (1979). *The structuring of organizations: a synthesis of the research*. Englewood Cliffs, NJ Prentice Hall International.

Mintzberg, H. (1993). *Structure in fives: designing effective organizations*. Englewood Cliffs, NJ Prentice Hall International.

Mintzberg, H. (2023). *Understanding organizations... Finally!: Structuring in sevens*. Englewood Cliffs, NJ Berrett-Koehler Publishers.

Misangyi, V. F., Greckhamer, T., Furnari, S., Fiss, P. C., Crilly, D., and Aguilera, R. V. (2017). Embracing causal complexity. *J. Manag.* 43, 255–282. doi: 10.1177/0149206316679252

Morrow, E., Zidaru, T., Ross, F., Mason, C., Patel, K. D., Ream, M., et al. (2023). Artificial intelligence technologies and compassion in healthcare: a systematic scoping review. *Front. Psychol.* 13:971044. doi: 10.3389/fpsyg.2022.971044

Nahavandi, S. (2019). Industry 5.0—a human-centric solution. *Sustainability* 11, 1–13. doi: 10.3390/su11164371

Nakao, Y., Strappelli, L., Stumpf, S., Naseer, A., Regoli, D., and Del Gamba, G. (2022). Towards responsible AI: a design space exploration of human-centered artificial intelligence user interfaces to investigate fairness. *Int. J. Hum. Comput. Interact.* 39, 1762–1788. doi: 10.1080/10447318.2022.2067936

Nisser, A., and Malanowski, N. (2019). Branchenanalyse chemische und pharmazeutische Industrie: Zukünftige Entwicklungen im Zuge Künstlicher Intelligenz. *Working Paper Forschungsförderung*, 166. Available at: http://hdl.handle.net/10419/216086

Organ, J. F., O'Neill, B. C., and Stapleton, L. (2021). Artificial intelligence and human-machine symbiosis in public employment services (PES): lessons from engineer and trade unionist, professor Michael Cooley. *IFAC-PapersOnLine* 387, 387–392. doi: 10.1016/j.ifacol.2021.10.478

Orlikowski, W. J. (1992). The duality of technology: rethinking the concept of technology in organizations. *Organ. Sci.* 3, 398–427. doi: 10.1287/orsc.3.3.398

Orlikowski, W. J. (2000). Using technology and constituting structures: a practice lens for studying technology in organizations. *Organ. Sci.* 11, 404–428. doi: 10.1287/orsc.11.4.404.14600

Orlikowski, W. J. (2007). Sociomaterial practices: exploring technology at work. *Organ. Stud.* 28, 1435–1448. doi: 10.1177/0170840607081138

Orlikowski, W. J., and Scott, S. V. (2008). 10 sociomateriality: challenging the separation of technology, work and organization. *Acad. Manag. Ann.* 2, 433–474. doi: 10.1080/19416520802211644

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). Updating guidance for reporting systematic reviews: development of the PRISMA 2020 statement. *J. Clin. Epidemiol.* 134, 103–112. doi: 10.1016/j.jclinepi.2021.02.003

Parker, S. K., and Grote, G. (2022). Automation, algorithms, and beyond: why work design matters more than ever in a digital world. *Appl. Psychol.* 71, 1171–1204. doi: 10.1111/apps.12241

Pava, C. (1986). Redesigning sociotechnical systems design: concepts and methods for the 1990s. *J. Appl. Behav. Sci.* 22, 201–221. doi: 10.1177/002188638602200303

Petropoulos, G. (2018). The impact of artificial intelligence on employment. *Praise Work Digital Age*, 119–132.

Plass, M., Kargl, M., Nitsche, P., Jungwirth, E., Holzinger, A., and Müller, H. (2022). Understanding and explaining diagnostic paths: toward augmented decision making. *IEEE Comput. Graph. Appl.* 42, 47–57. doi: 10.1109/mcg.2022.3197957

Ploug, T., and Holm, S. (2020). The four dimensions of contestable AI diagnostics - a patient-centric approach to explainable AI. *Artif. Intell. Med.* 107, 101901–101905. doi: 10.1016/j.artmed.2020.101901

Polak, S., Schiavo, G., and Zancanaro, M. (2022). Teachers' perspective on artificial intelligence education: an initial investigation," in Extended abstracts CHI conference on human factors in computing systems. 1–7

Riedl, M. O. (2019). Human-centered artificial intelligence and machine learning. *Hum. Behav. Emerg. Tech.* 1, 33–36. doi: 10.48550/arXiv.1901.11184

Riener, R., Frey, M., Bernhardt, M., Nef, T., and Colombo, G. (2005). "Human-centered rehabilitation robotics" in *2005 9th international conference on rehabilitation robotics (ICORR)* (Chicago, IL: IEEE), 319–322.

Riener, R., Lünenburger, L., and Colombo, G. (2006). Human-centered robotics applied to gait training and assessment. *J. Rehabil. Res. Dev.* 43, 679–694. doi: 10.1682/jrrd.2005.02.0046

Robert Bosch, GmbH (2020). *KI-Kodex von Bosch im Überblick*. Available at: https://assets.bosch.com/media/de/global/stories/ai_codex/bosch-code-of-ethics-for-ai.pdf

Romero, D., Stahre, J., Wuest, T., Noran, O., Bernus, P., Fast-Berglund, Å., et al. (2016). "Towards an operator 4.0 typology: a human-centric perspective on the fourth industrial revolution technologies" in *2016 46th international conference on computers and industrial engineering (CIE46)* (United States: Computers and Industrial Enfineering), 29–31.

Rožanec, J. M., Zajec, P., Kenda, K., Novalija, I., Fortuna, B., Mladenić, D., et al. (2021). STARdom: an architecture for trusted and secure human-centered manufacturing systems. *ArXiv*. doi: 10.1007/978-3-030-85910-7_21

Russell, S., and Norvig, P. (2021). *Artificial intelligence: a modern approach*. Upper Saddle River, NJ Prentice Hall.

Russo-Spena, T. R., Mele, C., and Marzullo, M. (2019). Practising value innovation through artificial intelligence: the IBM Watson case. *Journal of creating value* 5, 11–24. doi: 10.1177/2394964318805839

SAP SE (2021). *SAP's guiding principles for artificial intelligence*. Available at: https://www.sap.com/documents/2018/09/940c6047-1c7d-0010-87a3-c30de2ffd8ff.html

Schaal, S. (2007). The new robotics-towards human-centered machines. *HFSP journal* 1, 115–126. doi: 10.2976/1.2748612

Schmidtler, J., Knott, V., Hölzel, C., and Bengler, K. (2015). Human centered assistance applications for the working environment of the future. *Occupat. Ergon.* 12, 83–95. doi: 10.3233/OER-150226

Schönböck, J., Kurschl, W., Augstein, M., Altmann, J., Fraundorfer, J., Freller, L., et al. (2022). From remote-controlled excavators to digitized construction sites. *Proc. Comput. Sci.* 200, 1155–1164. doi: 10.1016/j.procs.2022.01.315

Seabra, D., Da Silva Santos, P. S., Paiva, J. S., Alves, J. F., Ramos, A., Cardoso, M. V. L. M. L., et al. (2022). The importance of design in the development of a portable and modular Iot-based detection device for clinical applications. *J. Phys.* 2292:012009. doi: 10.1088/1742-6596/2292/1/012009

Shaikh, S. J. (2020). Artificial intelligence and resource allocation in health care: the process-outcome divide in perspectives on moral decision-making. *AAAI Fall 2020 Symposium on AI for Social Good*. 1–8.

Shneiderman, B. (2020a). Bridging the gap between ethics and practice. *ACM Transact. Interact. Intellig. Syst.* 10, 1–31. doi: 10.1145/3419764

Shneiderman, B. (2020b). Human-centered artificial intelligence: reliable, safe & trustworthy. *Int. J. Hum. Comput. Interact.* 36, 495–504. doi: 10.1080/10447318.2020.1741118

Shneiderman, B. (2020c). Human-centered artificial intelligence: three fresh ideas. *AIS Transact. Hum. Comput. Interact.* 12, 109–124. doi: 10.17705/1thci.00131

Shneiderman, B. (2022). *Human-centered AI*. UK: Oxford University Press

Smith, N. L., Teerawanit, J., and Hamid, O. H. (2018). "AI-driven automation in a human-centered cyber world" in *2018 IEEE international conference of systems, man, and cybernetics (SMC)* (Miyazaki, JP: IEEE)

Snyder, H. (2019). Literature review as a research methodology: an overview and guidelines. *J. Bus. Res.* 104, 333–339. doi: 10.1016/j.jbusres.2019.07.039

Soldatos, J., and Kyriazis, D. (2021). *Trusted artificial intelligence in manufacturing: a review of the emerging wave of ethical and human centric AI technologies for smart production*. Boston: Now Publishers.

Springer, A. (2019). *Accurate, fair, and explainable: Building human-centered AI*. [dissertation]. [Santa Cruz]: University of California

Steels, L. (2020). "Personal dynamic memories are necessary to deal with meaning and understanding in human-centric AI" in *Proceedings of the first international workshop on new foundations for human-centered AI (NeHuAI)* (Santiago de Compostella, ES: CEUR-WS), 11–16.

Strohm, O., and Ulich, E. (1998). Integral analysis and evaluation of enterprises: a multi-level approach in terms of people, technology, and organization. *Hum. Fact. Ergon.*

Manufact. 8, 233–250. doi: 10.1002/(SICI)1520-6564(199822)8:3%3C233::AID-HFM3%3E3.0.CO;2-4

Suchman, L. (2012). "Configuration" in *Inventive methods. The happening of the social*. eds. C. Lury and N. Wakeford (London: Routledge), 48–60.

Suh, M., Youngblom, E., Terry, M., and Cai, C. J. (2021). AI as social glue: uncovering the roles of deep generative AI during social music composition. Proceedings of the 2021 CHI conference on human factors in computing systems, Yokohama, Japan 582, 1–11

Taryudi, T., Lindayani, L., Purnama, H., and Mutiar, A. (2022). Nurses' view towards the use of robotic during pandemic COVID-19 in Indonesia: a qualitative study. *Open Access Maced J. Med. Sci.* 10, 14–18. doi: 10.3889/oamjms.2022.7645

The White House (2022). Blueprint for an AI bill of rights. Making automated systems work for the American people. Available at: https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf

Tzachor, A., Devare, M., King, B., Avin, S., and Héigeartaigh, S. Ó. (2022). Responsible artificial intelligence in agriculture requires systemic understanding of risks and externalities. *Nat. Mach. Intellig.* 4, 104–109. doi: 10.1038/s42256-022-00440-4

Wang, D., Maes, P., Ren, X., Shneiderman, B., Shi, Y., and Wang, Q. (2021). "Designing AI to work WITH or FOR people?" in *Extended abstracts of the 2021 CHI conferende in human factors in computing systems* (New York, NY: Association for Computing Machinery)

Weekes, T. R., and Eskridge, T. C. (2022a). "Design thinking the human-AI experience of neurotechnology for knowledge workers" in *HCI international 2022 – Late breaking papers. Multimodality in advanced interaction environments*. Lecture Notes in Computer Science (Berlin: Springer Science + Business Media), 527–545.

Weekes, T. R., and Eskridge, T. C. (2022b). "Responsible human-centered artificial intelligence for the cognitive enhancement of knowledge workers" in *Lecture notes in computer science* (Berlin: Springer Science + Business Media), 568–582.

Wei, X., Ruan, M., Vadivel, T., and Daniel, J. A. (2022). Human-centered applications in sustainable smart city development: a qualitative survey. *J. Interconnect. Netw.* 22:2146001. doi: 10.1142/S0219265921460014

Widder, D. G., and Nafus, D. (2023). Dislocated accountabilities in the "AI supply chain": modularity and developers' notions of responsibility. *Big Data Soc.* 10. doi: 10.1177/20539517231177620

Wilkens, U. (2020). Artificial intelligence in the workplace – a double-edged sword. *Int. J. Informat. Lear. Technol.* 37, 253–265. doi: 10.1108/IJILT-02-2020-0022

Wilkens, U., Cost Reyes, C., Treude, T., and Kluge, A. (2021a). Understandings and perspectives of human-centered AI–A transdisciplinary literature review. *GfA Frühjahrskongress, B.10.17*.

Wilkens, U., Langholf, V., Ontrup, G., and Kluge, A. (2021b). "Towards a maturity model of human-centered AI–A reference for AI implementation at the workplace" in *Competence development and learning assistance systems for the data-driven future*. eds. W. Sihn and S. Schlund (Berlin, GE: GITO-Verlag), 179–197.

Wilkens, U., Lins, D., Prinz, C., and Kuhlenkötter, B. (2019). "Lernen und Kompetenzentwicklung in Arbeitssystemen mit künstlicher Intelligenz" in *Digitale transformation. gutes arbeiten und qualifizierung aktiv gestalten*. eds. D. Spath and B. Spanner-Ulmer (Berlin, GE: GITO-Verlag), 71–88.

Wilson, H. J., and Daugherty, P. R. (2018). Collaborative intelligence: humans and AI are joining forces. *Harv. Bus. Rev.* 96, 114–123.

Xie, Y., Gao, G., and Chen, X. (2019). "Outlining the design space of explainable intelligent systems for medical diagnosis" in *Joint proceedings of the ACM IUI workshops* (Los Angeles, USA: CEUR-WS)

Xu, W. (2019). Toward human-centered AI: a perspective from human-computer interactions. *Interactions* 26, 42–46. doi: 10.1145/3328485

Zhang, Y., Bellamy, R., and Varshney, K. (2020). "Joint optimization of AI fairness and utility: a human-centered approach" in *Proceedings of the AAAI/ACM conference on AI, ethics, and society* (New York, NY: ACM), 400–406.

Zhu, J., Liapis, A., Risi, S., Bidarra, R., and Youngblood, G. M. (2018). "Explainable AI for designers: a human-centered perspective on mixed-initiative co-creation" in *2018 IEEE conference on computational intelligence and games (CIG)* (Piscataway: IEEE), 1–8.

frontiers | Frontiers in Artificial Intelligence

TYPE Mini Review
PUBLISHED 03 November 2023
DOI 10.3389/frai.2023.1247755

OPEN ACCESS

EDITED BY
Annette Kluge,
Ruhr University Bochum, Germany

REVIEWED BY
Marie Ritter,
Technical University of Braunschweig, Germany
Simon Schwerd,
Universität der Bundeswehr München, Germany
Laura Kunold,
Ruhr University Bochum, Germany

*CORRESPONDENCE
Franziska Bocklisch
✉ franziska.bocklisch@mb.tu-chemnitz.de

RECEIVED 26 June 2023
ACCEPTED 10 October 2023
PUBLISHED 03 November 2023

CITATION
Bocklisch F and Huchler N (2023) Humans and cyber-physical systems as teammates? Characteristics and applicability of the human-machine-teaming concept in intelligent manufacturing.
Front. Artif. Intell. 6:1247755.
doi: 10.3389/frai.2023.1247755

Humans and cyber-physical systems as teammates? Characteristics and applicability of the human-machine-teaming concept in intelligent manufacturing

Franziska Bocklisch[1,2]* and Norbert Huchler[3]

[1]Department of Mechanical Engineering, Chemnitz University of Technology, Chemnitz, Germany, [2]Fraunhofer Institute for Machine Tools and Forming Technology, Chemnitz, Germany, [3]Institute for Social Science Research, Munich, Germany

The paper explores and comments on the theoretical concept of human-machine-teaming in intelligent manufacturing. Industrial production is an important area of work applications and should be developed toward a more anthropocentric Industry 4.0/5.0. Teaming is used a design metaphor for human-centered integration of workers and complex cyber-physical-production systems using artificial intelligence. Concrete algorithmic solutions for technical processes should be based on theoretical concepts. A combination of literature scoping review and commentary was used to identify key characteristics for teaming applicable to the work environment addressed. From the body of literature, five criteria were selected and commented on. Two characteristics seemed particularly promising to guide the development of human-centered artificial intelligence and create tangible benefits in the mid-term: complementarity and shared knowledge/goals. These criteria are outlined with two industrial examples: human-robot-collaboration in assembly and intelligent decision support in thermal spraying. The main objective of the paper is to contribute to the discourse on human-centered artificial intelligence by exploring the theoretical concept of human-machine-teaming from a human-oriented perspective. Future research should focus on the empirical implementation and evaluation of teaming characteristics from different transdisciplinary viewpoints.

KEYWORDS

human-machine-teaming, human-centered artificial intelligence, cognitive engineering, complementarity, shared knowledge and goals, human-centered industry 4.0/5.0

1. Introduction

1.1. Paper objectives

The technological evolution toward anthropocentric digitalization at work is rendered possible by new information and communication technologies as well as Artificial Intelligence (AI). It raises the questions: why and where is human-centered AI (HCAI) needed at work? Which recent theoretical concepts and methods can be applied to guide this complex, transdisciplinary endeavor in a responsible way? One good starting point is to clarify what "human-centeredness" means. As this is a very important but also general question, we use it as orientation to identify key characteristics and factors related to the

more focused concept of human-machine-teaming (HMT) and apply it to the working field of intelligent manufacturing. HMT can be defined as (1) a form of teamwork between humans and technical systems characterized by "real" interdependency between teammates such as joint activities toward a common goal (Johnson and Bradshaw, 2021). From another – more technical point of view – HMT may be characterized as (2) "the dynamic arrangement of humans and cyber-physical elements into a team structure that capitalizes on the respective strengths of each while circumventing their respective limitations in pursuit of shared goals" (Madni and Madni, 2018; p. 5). As these different transdisciplinary viewpoints on HMT may not be harmonized within one definition, we aim to capture key characteristics and criteria of HMT instead, using a literature review based on scoping method. The identified HMT criteria candidates are discussed and shortly illustrated by two example technologies from the working field of industrial manufacturing (human-robot-collaboration in assembly and intelligent decision support in thermal spraying). Our main objective is to contribute to the discourse on HCAI at work and to advance the development of the transdisciplinary, theoretical concept of HMT. Our comments come from a human-oriented perspective building on the research backgrounds from cognitive and engineering psychology as well as sociology of work and technology.

1.2. Human-centered artificial intelligence in industry

Generally, HCAI can be of interest in all areas of work in which complex problems have to be solved and a high level of security, speed, quality or efficiency of human-machine interactions is required. Among the fields are, for instance, military, medicine, mobility, finance, management and administrative knowledge work as well as intelligent manufacturing. The manufacturing industry is one of the most important economic sectors in the industrialized nations with a very high number of employees in various fields of work. The necessity of an anthropocentric perspective within Industry 4.0 is clearly recognized (see Rauch et al., 2020; Eich et al., 2023) and Xu et al. (2021) characterize the next step toward Industry 5.0 with its core values sustainability, resilience and true human-centeredness. Upcoming concepts such as human-cyber-physical systems (HCPS) show, how human-centeredness can be implemented concretely (Lamnabhi-Lagarrigue et al., 2017; Madni and Madni, 2018; Zhou et al., 2019; Bocklisch et al., 2022). HCPS combine three very different system parts: The human (H) in its two roles as user and developer of the technical system. The technical systems consists of (1) the physical subpart (P) controlled by (2) a cyber-system (C). Due to the complexity of manufacturing technologies and production processes, the C-part may implement AI algorithms. They represent effective means for machine control and should be developed toward HCAI (Shneiderman, 2022) and explainable AI (Hagras, 2018; Samek and Müller, 2019) to enable more joint working with humans and suitable support for cognitively demanding working tasks. Keep the human in the loop, is not primarily only a normative demand, but it is argued why this is functional (Huchler, 2022). Thus, humans have a special role in managing complexity in CPS (Böhle and Huchler, 2016). To that end HCPS offers a systemic and transdisciplinary perspective on automation allowing for flexibility and the development of semi-autonomous systems (Madni and Madni, 2018; Bocklisch et al., 2022). As a variety of industrial applications does not comply with the requirements for full automation and, furthermore, agility as well as (social) sustainability became increasingly important facets of modern work, the traditional, linear conceptualization of automation is not expedient. Hence, theoretical concepts for HCAI need to be derived from systemic and maybe even circular socio-technical concepts because (1) the technical developments effect use (and usefulness) of technical systems and the use (or misuse and disuse) has consequences for further developments and (2) automated systems are embedded again in social circumstances such as communication interfaces and work processes (Huchler, 2022). Circular concepts explicitly take into account the emergence of new forms of work or working tasks, being constantly created by automation of processes, systems and system components in various stages of technical development and use. In order to keep the human operator in the loop and combine human strengths with CP-systems capabilities in a complementary way, technical parts and AI algorithms should be developed in close accordance with human objectives and needs. Interests, discourses and narratives of the future drive technological innovations. They are subject to social dynamics between technology promises and disappointments, technological path dependencies, and changing images of man and technology. Recently, "human-centeredness" started to guide AI developments. Depending on the definition of AI used by the developers, the "similarity principle" may address cognitive aspects (e.g., models approximate human thinking or decision-making processes) or behavioral aspects (e.g., the final decision and intelligent machine behavior). Furthermore, the "difference principle" can mean that AI is "more rational than human cognition and behavior" (rational thought/action; cf. Russell, 2010, p. 2). If these different viewpoints in AI definitions are not payed attention to, one may easily misinterpret human-centeredness only as "similar" to the way, humans think, feel or act. However, true human-centeredness arises in the field of tension between the developmental opposites similarity (e.g., constituted by shared knowledge and shared goals; see application example 2.3.2 below) and difference/diversity (e.g., complementarity, non-redundant functions; see 2.3.1). Furthermore, human-centeredness may take different design metaphors as basis for AI and technological developments (cf. Figure 1, inner rectangle). For instance, AI may act as "supertool" or "tele-bot" vs. "intelligent agent" or "teammate" (Shneiderman, 2022). With regard to the chosen work application, we focus here on HMT because this concept may create tangible advantages and foster responsible solutions for industry in the mid-future. Compared to classical automation HMT is a rather transdisciplinary research field, that aims at integrating human-centered aspects into technology development more explicitly. This is done not only on a user-centered design level, but also more deeply, for instance, in the support or automation of cognitive processes (cf. example in Section 2.3.2; Bocklisch et al., 2022). This leads to a shift in goals: the goal of classical automation is to replace the human worker if possible. HMT aims at forming a joint work system with human and cyber-physical parts based on HCAI. It integrates

the potentials of both in new productive ways (Huchler, 2022) and may include a high degree of technical automation and human control (cf. Shneiderman, 2022). In the following, we review the concept of HMT with emphasis on finding key characteristics. Thereafter, we discuss the potential of two HMT criteria candidates for two industrial applications: human-robot-collaboration and intelligent decision-support. Other criteria are also reported and commented on. Then, we summarize which ones are (not yet) applicable and ready to be transferred from human-human-teams to human-cyber-physical-teams. Finally, we conclude and summarize future prospects for the HMT discourse and development.

2. Human-machine-teaming

HMT aims to transfer characteristics and principles of successful human-human-teams to human-cyber-physical-teams. This raises the question which features (= key characteristics) are ready and worth being implemented by HCAI in HCPS in the working field of production. Based on this, research can be planned into suitable methods and AI algorithms able to implement the identified features in the C-part.

2.1. Method

A structured literature review was performed starting with a scoping procedure (e.g., Arksey and O'Malley, 2005) to identify the breadth of contributions in HMT followed by a focused in-depth evaluation of records that present key characteristics of HMT for intelligent manufacturing. We understand key properties to be fundamental features of the theoretical HMT concept that may be addressed or implemented in some way in HCAI technology development in industrial applications in the near or mid-term future. The single keyword was "human-machine-teaming" and research results were limited to English documents between January 1 2016 and 31 May 2023 (no entries before 2016). For identification, the following databases revealed numerous records: scopus (N = 102) and Google scholar (N = 956). Exclusion and eligibility criteria were deliberately chosen rigorous in the second review phase. It was not the objective of this mini review to exhaustively review the research field of HMT or of related concepts (for this see Damacharla et al., 2018; O'Neill et al., 2022; Greenberg and Marble, 2023). Instead, we aimed to find key characteristics of HMT with sufficient conceptual strengths and high applicability to manufacturing that have already been taken up to a certain extend by the scientific community, to discuss them in-depth in terms of content (see 2.2) and illustrate them with the help of technological examples (see 2.3). After exclusion of redundant records, for 948 documents titles/abstracts were screened to identify eligibility (criterion was HMT definition by key characteristics) for full-text review (remaining N = 16 documents). After full text review, the remaining results were selected because they represent groundwork papers (N = 3: Brill et al., 2018; Madni and Madni, 2018; Johnson and Bradshaw, 2021). The HMT characteristics mentioned therein are discussed subsequently in the light of HCAI and industrial work context mainly from a cognitive psychology/human factors and work sociological point of view.

2.2. Selected key characteristics of human-machine-teaming

According to Madni and Madni (2018), HMT is the dynamic arrangement of humans and CPS into a team structure in pursuit of shared goals. Johnson and Bradshaw (2021) emphasize the interdependence relationship between teammates and point out that a team partner's behavior should be observable, predictable and directable. Brill et al. (2018) summarize the following facets for HMT: (1) complementarity, (2) shared knowledge and shared goals, (3) bounded autonomy, (4) mutual trust and (5) benevolence. Complementarity and shared knowledge/goals are related to how people make sense of situations in the field of tension between difference and similarity (Kelly, 1955). Therefore, these fundamental drivers also influence technical developments (e.g., difference: non-redundant complementary functions of technology compared to human capabilities vs. similarity: representation of human knowledge and goals in technical systems; see Figure 1, left). A meaningful sequence of development of HMT starts with these two criteria. Thereafter, the degree of automation or bounded autonomy of the cyber-part can be increased (see Figure 1, right; third criterion). Human trust in automation results from the transparent and successful implementation of these three characteristics. "Mutual trust" and "benevolence" are not applicable for manufacturing working applications (see Discussion). In the following, we focus on complementarity and shared knowledge/goals (see below) as those facets are already subject of HCAI-oriented research and at least – partly – studied in the context of manufacturing applications. Furthermore, they are prerequisites for bounded/semi-autonomy (Madni and Madni, 2018) and, hence, especially promising to establish a teaming relation.

2.3. Relevant aspects of human-machine-teaming in industrial working applications

Two aspects of HMT seem to be of special interest for industrial working applications: complementarity and shared knowledge/shared goals. With the help of two examples – one embodied and one un-embodied, cognitive technology – we outline the potential of these criteria in more detail.

2.3.1. Complementarity in human-robot-interaction

It is quite simple: two people who are able to accomplish the same working task may nevertheless share work and form a team. When a robot can do the same thing as a human team partner this usually results in full automation. Even better, in terms of flexibility and robustness of teamwork, is the combination of partners' abilities that complement each other (Huchler, 2020)

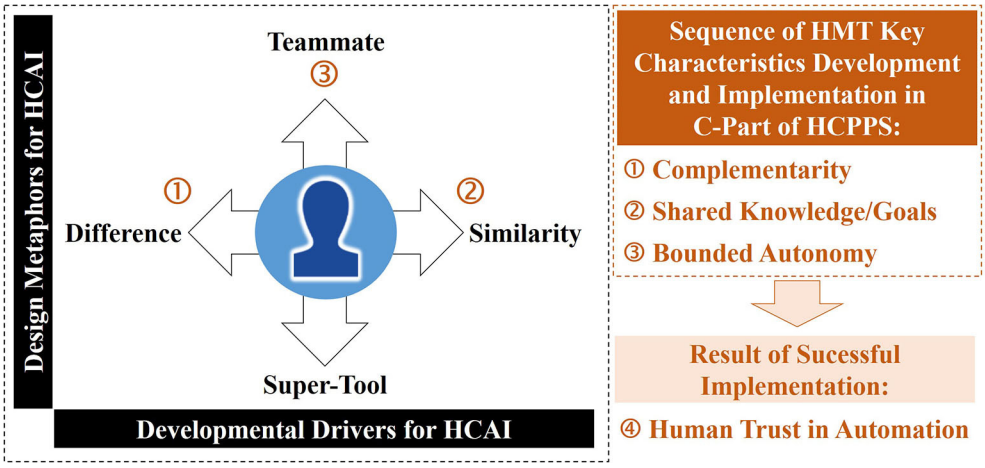

FIGURE 1
Human-centeredness as resulting balance between different technical design metaphors (**left**; vertical axis) and developmental drivers for Human-Centered Artificial Intelligence (HCAI; horizontal axis). Sequence of Human-Machine-Teaming (HMT) key characteristics development **(right)**. The first three characteristics are especially promising for industrial applications and should be integrated using HCAI in Human-Cyber-Physical-Production-Systems (HCPPS).

and may as well combine non-redundant strengths (Madni and Madni, 2018). Nevertheless, it is favorable if workers and robots have overlaps in their skills in a "mixed skill zone." This allows for adaptive interaction and may be organized in an AI-based human-centered way (Albu-Schäffer et al., 2023). The more humans and robots complement each other, the more productive interaction works (Huchler, 2022) affecting individual motivation at work in a positive way, for example, toward more effectiveness, empowerment, pride of production ("Produzentenstolz") and technology appropriation. Consequently, this increases trust in and social attachment to work tools in the second step. Similar to how construction workers feel enabled by an excavator in such a way that they "name" and maybe even "pet" it, collaborative robots can empower their human teammates as well. This feeling of support is based on complementarity and just not on similarity. Building on an extensive research line in industrial sociology on the particular relevance of work action and experiential knowledge in technologized work environments (e.g., Böhle and Milkau, 1988; Pfeiffer, 2007), Huchler et al. (2021) reported results of an extensive study in which the development and deployment process of an innovative robotic system for automated wiring of control cabinets was accompanied over 3 years (Huchler et al., 2021). The technical design approach initially chosen was mimicking the way humans work. It systematically narrowed developmental paths guiding directly toward the objective of full automation. The resulting technical solution was ineffective due to overwhelming complexity and automation limitations. A major problem was that there was no idea for productive worker involvement. As a result, the workers had to wait and repeatedly step in when the robot made mistakes. Furthermore, skill degradation, lack of integration of existing competencies as well as problems with allocation of functions and deployment were observed. The fallback solution after several attempts of correction was the complementary consideration of workers' cognitive and manual competencies resulting in the idea of a "supertool" workplace. The promise of cost savings through robotization was no longer linked to the simple idea of saving labor costs (substituting automation), but to increasing the productivity of existing employees (complementary automation). As a prerequisite for successful support in complex socio-technical contexts and HCPS, the places where people with their specific competencies are needed must be identified. Then socially sustainable and complementary HMT can be established. In this context, it is important to design the interaction as well as the permanent technological transformation in a "co-evolutionary" way so that people and technology can further develop along their different potentials in order to permanently create new complementarity relationships and maintain innovation capabilities (Huchler, 2022). These findings are supported by further qualitative and quantitative research on the relationship between human work capacities and collaborative lightweight robots (e.g., Pfeiffer, 2016, 2018).

2.3.2. Shared knowledge and goals in intelligent decision support for manufacturing

In manufacturing technologies needed for production of daily life goods, humans operate highly complex machines and technical processes such as in forming, welding or coating. Many technologies rely heavily on human expert knowledge and skills and, hence, can and will not be automated completely in the next future. Physical interactions have been improved by safety standards, worker protection and external means such as exoskeletons or use of robots (see above). However, due to technological and AI developments, system complexity increased rapidly shifting loads toward cognitive aspects (Darnstaedt et al., 2022). Hence, operators would benefit from cognitive augmentation and intelligent support for decision-making,

problem solving or fault diagnosis. A prerequisite for establishing a connection between a CPS and a human that resembles a human-human team relationship is that the team partners have a common understanding about the shared work task and goals. To achieve this, the knowledge representation in the CPS must be closely aligned with human expert knowledge (cf. Figure 1: similarity principle) to enable transparent understanding and good interactions. Otherwise, there is a risk that the CPS will represent something (e.g., from sensor data) that has no substantive meaning for humans. If this is the case, then there is no good basis for human-centered and joint teamwork, for example, joint decision-making in complex situations. This research gap is recognized and partly addressed with AI for different manufacturing technologies such as coating (Bobzin et al., 2022; Mahendru et al., 2023). These solid domain-oriented research approaches should be enriched by focusing more explicitly on the human perspective. For instance, by considering action-guiding rules for optimization of technical parameters (Venkatachalapathy et al., 2023) or elicitation of domain knowledge and expert mental models (Hoffman, 2008; Andrews et al., 2023). Sharing knowledge and goals in the sense of how a human "shares" ideas with another human is challenging. First, relevant knowledge needs to be elicited. This is possible but only within the boundaries of what can be brought to consciousness (expert-driven approach; Hoffman et al., 2021) or what can be measured and interpreted semantically without doubt (data-driven approach). Nevertheless, it will never be "complete" compared to the human treasure trove of experience, which is continuously growing and can only be described and formalized in parts (Huchler, 2017). Second, the elicited knowledge requires transparent and strictly HCAI to form an interdependence relationship that is mutually explain- and understandable. In order to do so, a combination of different AI algorithms – knowledge- and data-based methods – are needed to ensure compatibility with different human performance levels such as skill-, rule- or knowledge-based behavior (Rasmussen, 1983). Pure sensory- and data-based procedures will not form a sufficient basis for HMT the intelligent manufacturing because they can only grasp a limited area of what is actually necessary (Rasmussen, 1983; Bocklisch and Lampke, 2023; mainly skill-based behavior).

3. Discussion

3.1. Key characteristics of human-machine-teaming in industrial working applications

HMT is an innovative concept with potential for real-world working domains such as manufacturing. It may guide HCAI developments toward more anthropocentric designs, new forms of work and human-machine interaction. Based on a review of recent literature as well as own preliminary work, we consider the systematization of Brill et al. (2018) as one good starting point for in-depth discussion of potential teaming characteristics for HCAI in industrial manufacturing. In Figure 1, the criteria have been systematized and placed in a meaningful order of development and implementation in HCPPS. Criteria "complementarity" and "shared knowledge/goals" have been illustrated with concrete examples (see above), because (a) they have already been researched to a certain extent in the work context of intelligent manufacturing and (b) they represent essential foundations for criteria "bounded autonomy" and "trust." In the following, the criteria are discussed in detail, placed in an overall context, and illuminated with regard to future research needs.

(1) Complementarity: yes, in our opinion this criterion is central for HMT because the dissimilarity/diversity facet and may be used to augment humans by powerful complementary functionalities that are provided by the cyber-physical-production-system (CPPS). However, this is not a static concept but characterized as ongoing innovation process – including permanent search for new potentials for complementarity and (re)adjustment of education and further training. Hence, there is need for a better understanding of the differences of human and technology/AI as well as of automation dynamics and changes in the human-technology relationship.

(2) Shared knowledge/goals: These criteria refer to the opposite of complementarity and use similarity principle to constitute a common working basis between humans and CP-systems. A successful and reliable working relation as well as efficient function allocations need shared knowledge and goals. Both, implicit and explicit forms of human knowledge are needed in working contexts. Hence, cognitive engineering methods for knowledge elicitation, structured systematization and transparent AI-implementation need to be developed further. Joint goals can potentially be defined on various levels of abstraction. High-level experts, for instance, persons controlling complex plants, are able to use their rich knowledge hierarchies and related procedures to tackle concrete situations in a very flexible way (Rasmussen, 1983). Changes in the situation are managed by goal or sub goal adaptation. These human strategies to control real-world complexity and act under uncertainty need to be mirrored – at least partly – in the cyber-teammate as well. If this can be achieved successfully will depend on the development of AI regarding adaptivity and learning (e.g., evolving intelligent systems: Angelov et al., 2010; Bocklisch et al., 2017) as well as cognitive transparency and understandability of AI algorithms (e.g., Weller, 2019).

(3) Bounded autonomy: autonomy is always limited and negotiated in social contexts. For HMT, different kinds of autonomies have to be integrated similar to the different "intelligences" (human vs. artificial). The simple technical levels of autonomy (e.g., functionality within a limited context) do not correspond to the complexity of the socially negotiated understanding of autonomy of individuals. As with intelligence, the complexity of the social counterpart is completely underestimated or taken too simplistically. Hence, profound conceptual research should relate theoretical concepts to concrete application examples. This is also necessary because autonomy is a "provocative" criterion that may easily lead to conflicting viewpoints (Brill et al., 2018) as well as fears from the human user side. Technology assessments that evaluate dangers (see "The janus face of autonomy" in Brill et al., 2018) as well as possibilities and derive regulatory principles (Shneiderman, 2020) are therefore needed as well.

(4) Mutual trust and (5) benevolence: Trust is central to establish a successful and harmonic relationship in human-human work teams. One classic definition originates from Lee and See (2004; p. 54): trust is "… the attitude that an agent will help achieve

an individual's goals in a situation characterized by uncertainty and vulnerability." In this respect, it is a good candidate criterion worth being thought of concerning its transferability to HM-teams and closely related to "shared goals" – a part of the definition and thus a necessary condition for trust. Trust in automation is extensively studied (e.g., Lee and Moray, 1992; Hoff and Bashir, 2015; Schaefer et al., 2016; Kohn et al., 2021) and a highly important factor for user-centered design to avoid misuse, disuse or abuse of technology (Parasuraman and Riley, 1997; Lee and See, 2004). Nevertheless, "trust is a complex and nebulous concept" (Hoffman et al., 2013, p. 84) and should not be understood in a too simplistic way as a "lack of information" but rather as a complex process of (reciprocally effective!) establishing the ability to act even beyond (risk) calculations (Huchler and Sauer, 2015). Furthermore, it seems only applicable from a human point of view: a human trusts a robotic system or a suggestion of a decision support system (more precisely: the people and institutions behind). The relation cannot simply be reversed and named "trust" because trust presupposes physical and/or mental vulnerability, which applies to technology only to a very limited extent. Sociological aspects are important to consider as well. What is often perceived as "trustful relationship" to a technical artifact (similar to a person) is in reality based on social processes (Mayer et al., 1995) in a complex social-technical setting primarily also related to trust in the institutions responsible for technology. This explains some experimental results concerning "over trust" in robots (Aroyo et al., 2021). The institutions and regularities are important guarantors for safety. At least in work contexts, it is evident that trust in and acceptance of technology can be generated much more clearly through utility and empowerment than through similarity which is only one of the polar development drivers (cf. Figure 1). From the human user perspective, too close similarity to human skills comes with a latent threat: substitutability – the opposite of benevolence, which is in our opinion no primary target criterion for HMT. "Mutual" trust and benevolence are no purposeful facets for HMT because technology is not able to trust or act benevolent. Here, the distinction between system trust and personal trust is crucial (Luhmann, 1979). Nevertheless, suitable objective criteria from the technical point of view have to be developed instead.

3.2. Limitations and future prospects

Our main objective was to contribute to the discourse on HCAI by having a closer look on the theoretical concept of HMT in the context of industrial work applications. This is intended to be an impulse from a human-oriented perspective on AI developments for future transdisciplinary discourses. Of course, there are many other perspectives on this topic that are equally interesting, relevant and necessary. For example, concepts and empirical work from research on human teamwork (e.g., concerning suitable definitions of "team" and types of teams) and team performance as well as from (software) engineering are crucial for complementing and validating HMT criteria. Here, our focus was on theoretical considerations but guides on the implementation of HMT aspects already exist, highlighting the practical relevance of the topic (e.g., McDermott et al., 2018). Industry 4.0/5.0 developers would benefit from operationalizing various HMT criteria in industrial examples. Not only on the general level of user-centered design guides but more in-depth for specific technical applications (Bocklisch et al., 2022). Another limitation was the narrow scope of search terms: given the huge number of literature and our specific goal to find applicable key characteristics for manufacturing and comment them in the light of two short application examples, we only selected "HMT" as keyword for scoping review. Other words, such as "human-autonomy-teaming," "human-agent-teaming," "human-machine-interaction," "human-machine-symbiosis," and many thematically related terms in various combinations would lead to a more comprehensive and – concerning the vast body of empirical evidence – less biased summary (cf. O'Neill et al., 2022). Furthermore, we did not discuss all potential HMT-criteria as key features but reduced to five aspects from which we selected two to outline their concrete potential for industrial applications with the help of two technical examples. On the one hand, this specific procedure and scope resulted from the fact that some facets clearly need to be given ex ante to be of interest for HCAI (such as observability; cf. 2.3.2 and boundaries of human knowledge elicitation and data acquisition from human sources). On the other hand, this was because some criteria are very similar and somehow eclectic (e.g., bounded autonomy vs. semi-autonomy or interdependency). Whether these slightly different connotations of criteria, e.g., of the core characteristic "bounded autonomy," should be taken into account cannot be adequately assessed at present. This will be shown by the operationalization of the characteristics in the empirical work, the practical application and the evaluation of these results.

In conclusion, HCAI has a large potential to promote new types of human-machine-interaction at work, such as outlined here in parts for HMT. The transfer of some characteristics of HH-teams to HCP-teams are promising and feasible for real-world working contexts such as intelligent manufacturing, others not – because humans and technology are very different in nature (Madni and Madni, 2018; p. 4f) – or not yet – because HCAI capabilities still need to be developed further. If HMT capabilities are to be integrated into technology development of HCPS as a concrete form of HCAI, then the start could – in our opinion – be to establish complementarity and shared knowledge/goals. Thereafter, the effects of this development should be evaluated from different viewpoints that are important in intelligent manufacturing such as human-oriented criteria (e.g., user acceptance, mental workload), technical or business oriented aspects (e.g., system performance, product quality, resource efficiency and costs).

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

The publication of this article was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project number 491193532 and the Chemnitz University of Technology. This work was supported by the Fraunhofer internal programs under grant: Attract 40-06107.

Acknowledgments

We thank Thomas Lampke, Marcel Todtermuschke, and Steffen Bocklisch for discussions about human-machine teaming concepts from a technical point of view and three reviewers for their valuable feedback that helped to improve the paper.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Albu-Schäffer, A., Huchler, N., Kessler, I., Lay, F., Perzylo, A., Seidler, M., et al. (2023). Soziotechnisches assistenzsystem zur lernförderlichen arbeitsgestaltung in der robotergestützten montage. Gruppe interaktion organisation. *Zeitschrift Angew. Org.* 54, 79–93. doi: 10.1007/s11612-023-00668-7

Andrews, R. W., Lilly, J. M., Srivastava, D., and Feigh, K. M. (2023). The role of shared mental models in human-AI teams: a theoretical review. *Theor. Issues Erg. Sci.*, 24, 129–175. doi: 10.1080/1463922X.2022.2061080

Angelov, P., Filev, D. P., and Kasabov, N. (2010). *Evolving intelligent Systems: Methodology and Applications.* London: John Wiley and Sons.

Arksey, H., and O'Malley, L. (2005). Scoping studies: towards a methodological framework. *Int. J. Soc. Res. Methodol.* 8, 19–32. doi: 10.1080/1364557032000119616

Aroyo, A. M., De Bruyne, H., Dheu, J., Fosch-Villaronga, O., Gudkov, E., Hoch, A., et al. (2021). Overtrusting robots: setting a research agenda to mitigate overtrust in automation. *Paladyn J. Behav. Robotic.* 12, 423–436. doi: 10.1515/pjbr-2021-0029

Bobzin, K., Heinemann, H., and Dokhanchi, S. R. (2022). Development of an expert system for prediction of deposition efficiency in plasma spraying. *J. Therm. Spray Technol.* 32, 643–656. doi: 10.1007/s11666-022-01494-x

Bocklisch, F., Bocklisch, S. F., Beggiato, M., and Krems, J. F. (2017). Adaptive fuzzy pattern classification for the online detection of driver lane change intention. *Neurocomputing* 262, 148–158. doi: 10.1016/j.neucom.2017.02.089

Bocklisch, F., and Lampke, T. (2023). *Mensch und Maschine als Teampartner? Neue Wege zur Menschzentrierten Digitalisierung in der Produktion*. Singapore: WOMAG.

Bocklisch, F., Paczkowski, G., Zimmermann, S., and Lampke, T. (2022). (2022). Integrating human cognition in cyber-physical systems: A multidimensional fuzzy pattern model with application to thermal spraying. *J. Manuf. Syst.* 63, 162–176. doi: 10.1016/j.jmsy.2022.03.005

Böhle, F., and Huchler, N. (2016). "Cyber-Physical Systems and Human Action. A re-definition of distributed agency between humans and technology, using the example of explicit and implicit knowledge," in *Cyber-Physical Systems: Foundations, Principles, and Applications. A volume in Intelligent Data-Centric Systems*, eds H. Song, D. B. Rawat, S. Jeschke, and C. Brecher (Elsevier), 115–127. doi: 10.1016/B978-0-12-803801-7.00008-0

Böhle, F., and Milkau, B. (1988). *Vom Handrad zum Bildschirm - Eine Untersuchung zur sinnlichen Erfahrung im Arbeitsprozeß. Campus.*

Brill, C. J., Cummings, M. L., Evans, I. I. I., Hancock, A. W., and Lyons, P. A. J. B., and Oden, K. (2018). Navigating the advent of human-machine teaming. *Proc. Human Factors Erg. Soc. Ann. Meeting.* 62, 455–459. doi: 10.1177/1541931218621104

Damacharla, P., Javaid, A. Y., Gallimore, J. J., and Devabhaktuni, V. K. (2018). Common metrics to benchmark human-machine teams (HMT): a review. *IEEE Acc.* 6, 38637–38655. doi: 10.1109/ACCESS.2018.2853560

Darnstaedt, D. A., Ahrens, A., Richter-Trummer, V., Todtermuschke, M., and Bocklisch, F. (2022). Procedure for describing human expert knowledge and cognitive processes during the teach-in of industrial robots. *Zeitschrift für Arbeitswissenschaft* 4, 1–16. doi: 10.1007/s41449-021-00284-5

Eich, A., Klichowicz, A., and Bocklisch, F. (2023). How automation level influences moral decisions of humans collaborating with industrial robots in different scenarios. *Front. Psychol.* 14, 1107306. doi: 10.3389/fpsyg.2023.1107306

Greenberg, A. M., and Marble, J. L. (2023). Foundational concepts in person-machine teaming. *Front. Phys.* 10, 1310. doi: 10.3389/fphy.2022.1080132

Hagras, H. (2018). Toward human-understandable, explainable AI. *Computer* 51, 28–36. doi: 10.1109/MC.2018.3620965

Hoff, K. A., and Bashir, M. (2015). Trust in automation: integrating empirical evidence on factors that influence trust. *Hum. Factors* 57, 407–434. doi: 10.1177/0018720814547570

Hoffman, R. R. (2008). Human factors contributions to knowledge elicitation. *Hum. Fact.* 50, 481–488. doi: 10.1518/001872008X288475

Hoffman, R. R., Johnson, M., Bradshaw, J. M., and Underbrink, A. (2013). Trust in automation. *IEEE Int. Syst.* 28, 84–88. doi: 10.1109/MIS.2013.24

Hoffman, R. R., Klein, G., and Mueller, S. T. (2021). A Guide to the Measurement and Evaluation of User Mental Models. Technical Report, DARPA Explainable AI Program.

Huchler, N. (2017). Grenzen der Digitalisierung von Arbeit – Die Nicht-Digitalisierbarkeit und Notwendigkeit impliziten Erfahrungswissens und informellen Handelns. *Z. Arbeitswissenschaft* 71, 215–223. doi: 10.1007/s41449-017-0076-5

Huchler, N. (2020). *Die Mensch-Maschine-Interaktion bei KI in der Arbeit Menschengerecht Gestalten? Das HAI-MMI Konzept und die Idee der Komplementaritat. Digitale Welt.* Available online at: https://digitaleweltmagazin.de/en/fachbeitrag/die-mensch-maschine-interaktion-bei-kuenstlicher-intelligenz-im-sinne-der-beschaeftigten-gestalten-das-hai-mmi-konzept-und-die-idee-der-komplementaritaet/ (accessed March 15, 2023).

Huchler, N. (2022). Komplementäre arbeitsgestaltung. grundrisse eines konzepts zur humanisierung der arbeit mit KI. *Zeitschrift für Arbeitswissenschaft* 76, 158–175. doi: 10.1007/s41449-022-00319-5

Huchler, N., Kessler, I., Lay, F. S., Perzylo, A., Seidler, M., Steinmetz, F., et al. (2021)."Empowering workers in a mixed skills concept for collaborative robot systems," in *Workshop on Accessibility of Robot Programming and Work of the Future, Robotics: Science and Systems (RSS 2021)*. Cologne: German Aerospace Center.

Huchler, N., and Sauer, S. (2015). Reflexive and experience-based trust and participatory research: concept and methods to meet complexity and uncertainty in organisations. *Int. J. Action Res.* 11, 146–173.

Johnson, M., and Bradshaw, J. M. (2021). "How interdependence explains the world of teamwork," in *Engineering Artificially Intelligent Systems: A Systems Engineering Approach to Realizing Synergistic Capabilities, LNCS,* eds W. F.Lawless (Cham: Springer), 122–146.

Kelly, G. A. (1955). *The Psychology of Personal Constructs: A Theory of Personality, Vol 1.* New York, NY: WW Norton and Company.

Kohn, S. C., De Visser, D, Wiese, E. J., Lee, E. Y. C., and Shaw, T. H. (2021). Measurement of trust in automation: a narrative review and reference guide. *Front. Psychol.* 12, 604977. doi: 10.3389/fpsyg.2021.604977

Lamnabhi-Lagarrigue, F., Annaswamy, A., Engell, S., Isaksson, A., Khargonekar, P., Murray, R. M., et al. (2017). Systems and control for the future of humanity, research agenda: current and future roles, impact and grand challenges. *Ann. Rev. Control* 43, 1–64. doi: 10.1016/j.arcontrol.2017.04.001

Lee, J., and Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 35, 1243–1270. doi: 10.1080/00140139208967392

Lee, J., and See, K. A. (2004). Trust in automation: designing for appropriate reliance. *Hum. Factors* 46, 50–80. doi: 10.1518/hfes.46.1.50.30392

Luhmann (1979). *Trust and Power: Two Works*. New York, NY: Wiley.

Madni, A. M., and Madni, C. C. (2018). Architectural framework for exploring adaptive human-machine teaming options in simulated dynamic environments. *Systems* 6, 44. doi: 10.3390/systems6040044

Mahendru, P., Tembely, M., and Dolatabadi, A. (2023). Artificial intelligence models for analyzing thermally sprayed functional coatings. *J. Therm. Spray Technol.* 32, 388–400. doi: 10.1007/s11666-023-01554-w

Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An integrative model of organizational trust. *Acad. Manage. Rev* 20, 709–734. doi: 10.2307/258792

McDermott, P., Dominguez, C., Kasdaglis, N., Ryan, M., Trahan, I., Nelson, A., et al. (2018). *Human-Machine Teaming Systems Engineering Guide*. Bedford, MA: MITRE Corp.

O'Neill, T., McNeese, N., Barron, A., and Schelble, B. (2022). Human–autonomy teaming: a review and analysis of the empirical literature. *Hum. Fact.* 64, 904–938. doi: 10.1177/0018720820960865

Parasuraman, R., and Riley, V. (1997). Humans and automation: use, misuse, disuse, abuse. *Hum. Fact.* 39, 230–253. doi: 10.1518/001872097778543886

Pfeiffer, S. (2007). *Montage und Erfahrung – Warum Ganzheitliche Produktionssysteme menschliches Arbeitsvermögen brauchen*. Verlag: Rainer Hamp Verlag.

Pfeiffer, S. (2016). Robots, industry 4.0 and humans, or why assembly work is more than routine work. *Societies* 6, 16. doi: 10.3390/soc6020016

Pfeiffer, S. (2018). Industry 4, 0. robotics and contradictions. *Technol. Lab. Polit. Contradic.* 12, 19–36. doi: 10.1007/978-3-319-76279-1_2

Rasmussen, J. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Trans. Syst. Man Cybernet.* 3, 257–266. doi: 10.1109/TSMC.1983.6313160

Rauch, E., Linder, C., and Dallasega, P. (2020). Anthropocentric perspective of production before and within Industry 4.0. *Comput. Ind. Eng.* 139, 105644. doi: 10.1016/j.cie.2019.01.018

Russell, S. J. (2010). *Artificial Intelligence a Modern Approach*. London: Pearson Education, Inc.

Samek, W., and Müller, K. R. (2019). "Towards explainable artificial intelligence," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning,* eds W. Samek and G Montavon (Cham: Springer International Publishing), 5–22.

Schaefer, K. E., Chen, J. Y., Szalma, J. L., and Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Hum. Fact.* 58, 377–400. doi: 10.1177/0018720816634228

Shneiderman, B. (2020). Human-centered artificial intelligence: three fresh ideas. *AIS Trans. Hum. Comp. Int.* 12, 109–124. doi: 10.17705/1thci.00131

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

Venkatachalapathy, V., Katiyar, N. K., Matthews, A., Endrino, J. L., and Goel, S. (2023). A guiding framework for process parameter optimisation of thermal spraying. *Coatings* 13, 713. doi: 10.3390/coatings13040713

Weller, A. (2019). "Transparency: motivations and challenges," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning,* eds W. Samek and G Montavon (Cham: Springer International Publishing), 23–40.

Xu, X., Lu, Y., Vogel-Heuser, B., and Wang, L. (2021). Industry 4.0 and industry 5.0—inception, conception and perception. *J. Manuf. Syst.* 61, 530–535. doi: 10.1016/j.jmsy.2021.10.006

Zhou, J., Zhou, Y., Wang, B., and Zang, J. (2019). Human-cyber-physical systems (HCPSs) in the context of new-generation intelligent manufacturing. *Engineering* 5, 624–636. doi: 10.1016/j.eng.2019.07.015

frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 17 November 2023
DOI 10.3389/frai.2023.1253940

OPEN ACCESS

EDITED BY
Corinna Peifer,
University of Lübeck, Germany

REVIEWED BY
Stefano Triberti,
Pegaso University, Italy
Michèle Rieth,
University of Bremen, Germany

*CORRESPONDENCE
Anna-Stiina Wallinheimo
anna-stiina.wallinheimo@surrey.ac.uk

†These authors have contributed equally to this work

RECEIVED 06 July 2023
ACCEPTED 19 October 2023
PUBLISHED 17 November 2023

CITATION
Wallinheimo A-S, Evans SL and Davitti E (2023) Training in new forms of human-AI interaction improves complex working memory and switching skills of language professionals. *Front. Artif. Intell.* 6:1253940. doi: 10.3389/frai.2023.1253940

Training in new forms of human-AI interaction improves complex working memory and switching skills of language professionals

Anna-Stiina Wallinheimo[1,2*†], Simon L. Evans[2†] and Elena Davitti[1†]

[1]Centre for Translation Studies, Faculty of Arts and Social Sciences (FASS), University of Surrey, Guildford, United Kingdom, [2]School of Psychology, Faculty of Health and Medical Sciences (FHMS), University of Surrey, Guildford, United Kingdom

AI-related technologies used in the language industry, including automatic speech recognition (ASR) and machine translation (MT), are designed to improve human efficiency. However, humans are still in the loop for accuracy and quality, creating a working environment based on Human-AI Interaction (HAII). Very little is known about these newly-created working environments and their effects on cognition. The present study focused on a novel practice, interlingual respeaking (IRSP), where real-time subtitles in another language are created through the interaction between a human and ASR software. To this end, we set up an experiment that included a purpose-made training course on IRSP over 5 weeks, investigating its effects on cognition, and focusing on executive functioning (EF) and working memory (WM). We compared the cognitive performance of 51 language professionals before and after the course. Our variables were reading span (a complex WM measure), switching skills, and sustained attention. IRSP training course improved complex WM and switching skills but not sustained attention. However, the participants were slower after the training, indicating increased vigilance with the sustained attention tasks. Finally, complex WM was confirmed as the primary competence in IRSP. The reasons and implications of these findings will be discussed.

KEYWORDS

AI-related technologies, automatic speech recognition (ASR), interlingual respeaking (IRSP), human-AI interaction (HAII), cognition, executive function (EF), working memory (WM)

1 Introduction

In the language industry, which is currently one of the fastest growing industries (CSA Research, 2023), AI-related technologies, including automatic speech recognition (ASR) and machine translation (MT), have been created to automate repetitive and time-pressured tasks. However, these technologies are currently not sufficiently accurate to be used alone: human input is needed for sense checking and quality control. Humans are, therefore, responsible for monitoring and possibly correcting the written output produced by technology through Human-AI Interaction (HAII). Thus, AI-related technologies intended to extend and improve human efficiency are increasing the number of tasks people deal with, leading to new cognitive environments for professionals in the language industry and presenting new cognitive challenges and requirements.

In this paper, which draws on a wider experiment carried out within the framework of the ESRC-funded SMART project (*Shaping Multilingual Access through Respeaking Technology*, ES/T002530/1, 2020–2023), we will focus on a novel practice that relies on HAII, namely interlingual respeaking (IRSP). In IRSP, real-time subtitles in another language are created through the interaction of a human and ASR software (Davitti and Sandrelli, 2020; Pöchhacker and Remael, 2020). IRSP is a cognitively demanding real-time process where a language professional simultaneously translates incoming spoken language while adding punctuation and content labels orally, as well as applying any editing needed to ASR, which turns what they say into subtitles (Davitti and Sandrelli, 2020; Pöchhacker and Remael, 2020). This is a multi-step process where humans and technology need to work together to be able to produce highly-accurate subtitles in a timely manner.

Since the early 2000s, respeaking has been widely employed to produce intralingual subtitles (i.e., in the same language) for d/Deaf and Hard-of-Hearing audiences (Romero-Fresco, 2011). The recent worldwide increase in audiovisual content has led to an ever-increasing demand for making this content accessible across languages and in real time—hence the rise of interlingual respeaking (i.e., from one language to another), which adds language transfer to the traditional respeaking practice.

Pöchhacker and Remael (2020) conducted a detailed theoretical analysis of the IRSP process to guide future studies into the competences and skills required to perform it. In this newly-created process and competence model, the cognitive resources required for the IRSP process are placed in the technical-methodological competence area (Pöchhacker and Remael, 2020). However, the required cognitive functions are based on a competence-oriented task analysis of the IRSP process rather than on experimental investigations. Thus, the current study aims to bring more depth and empirical evidence to these initial findings. To this end, we set up an experiment that included a purpose-made training course on IRSP. We investigated what cognitive resources measured prior to the training predicted higher IRSP accuracy post-training. As part of the investigation, we also explored how the training course affected human cognition, executive functions (EF) and working memory (WM), in particular. We were interested in these cognitive functions as distributed cognition (DCog) posits that integrating technological tools with internal cognitive resources can increase the mental workspace available (Kirsh, 1995; Wallinheimo et al., 2019). However, little is known about how HAII affects human cognition, particularly when applied to real-time practices involving multiple tasks conducted under severe time constraints (as required by IRSP). As HAII becomes more common in the fast-evolving modern workplace, knowledge around the links to an individual's cognitive processes is needed to allow for people-centered and responsible AI.

1.1 New cognitive environment based on DCog

IRSP creates a new cognitive environment where human cognition is distributed to the outside world by relying on technological tools i.e., the use of ASR. DCog aims to understand the organization of human cognitive systems by extending what is traditionally considered cognitive beyond an individual doing the task to include interactions between the people involved in the process and the external resources e.g., technological tools in the environment (Hutchins, 1995; Hollan et al., 2000). In DCog, a new broader unit of cognitive analysis is created that includes a network of technologies and actors leading to a process that coordinates internal processes in the minds of humans working together, with external representations relying on external artifacts. According to Vallée-Tourangeau and Vallée-Tourangeau (2017), thinking can be seen as a cognitive process that develops in time and space and leads to a new cognitive event, for example, a solution to a problem. These cognitive events emerge from cognitive interactivity, which the authors define as "the meshed network of reciprocal causations between an agent's mental processing and transformative actions she applies to her immediate environment to achieve a cognitive result" (Vallée-Tourangeau and Vallée-Tourangeau, 2017).

Thus, novel forms of HAII give rise to new working environments that impact existing cognitive processes in distinct ways. Several experimental studies have explored individual problem solving to examine cognitive interactivity, yielding valuable insights into the use and benefits of distributing human cognition, with a particular focus on the cognitive needs of the individual. Wallinheimo et al. (2019) found that when evaluative pressure is experienced to complete a cognitive task, there are additional demands on the existing limited WM resources. However, some of these WM limitations, caused by the additional worry of performing well, can be compensated by offloading the cognitive process to the external environment by using pen and paper or other external artifacts (Wallinheimo et al., 2019). This is in line with Risko and Gilbert (2016), who argue that cognitive offloading reduces the overall cognitive demand on the individual (Risko and Gilbert, 2016). Additionally, Kirsh (2010) claimed that there are cost-benefit considerations for cognitive interactivity, and as a result, cognitive processes go to wherever it is easier to perform them. It might be easier to understand a complicated sentence by drawing a picture of it, to visualize it externally rather than just thinking internally in the head alone. Thus, the overall cognitive cost of sense making to understand the sentence is reduced with the help of drawing a picture (Kirsh, 2010). Finally, cognitive interactivity can extend WM resources when there is a cognitive need to do so (e.g., less efficient WM capabilities) (Webb and Vallée-Tourangeau, 2009).

When it comes to IRSP, however, both the human and the machine have equally important roles to make the IRSP process work as both are required to work in synchrony for the creation of accurate interlingual subtitles. This is a form of human-autonomy teamwork (HAT) where humans work interdependently with an autonomous agent (i.e., ASR) focusing on a set of tasks toward the shared goal of producing interlingual live subtitles (O'Neill et al., 2022). Thus, during the IRSP process, human cognition is distributed with the use of ASR not due to a cognitive need of the language professional (i.e., reduced WM capacity, cost-benefit considerations, or cognitive offloading), but rather as a requirement inherent to the IRSP process itself. This leads to a distinct cognitive environment that sets it apart from the experimental studies mentioned earlier.

1.2 IRSP and simultaneous interpreting

IRSP is a new practice, and empirical investigations into the cognitive processes involved are in their infancy. As a real-time language-related practice, IRSP shares many aspects with simultaneous interpreting (SI), which is widely acknowledged as one of the most cognitively challenging tasks of human cognition (Babcock and Vallesi, 2017). Hence, we have drawn upon the existing SI literature as the starting point of our investigation, recognizing its relevance in understanding the cognitive intricacies of IRSP. When simultaneously interpreting, an interpreter needs to concurrently comprehend auditory material in one language while producing the same content in another language. Executive functions (EF) are heavily involved in this process. EFs are a set of cognitive processes that are needed for the cognitive control of human behavior. The three most postulated areas of EF are: shifting between tasks or mental sets (shifting), information updating and monitoring in WM (updating), and inhibition of prepotent responses (inhibition) (Miyake et al., 2000). SI requires both short-term memory and WM resources to keep the required information active and to be able to manipulate it throughout the SI process (Timarova, 2007; Aben et al., 2012; Mellinger and Hanson, 2019). Additionally, for the simultaneous interpreter to keep control of incoming information and avoid mixing languages, effective recall and attentional and cognitive control are needed (Christoffels and de Groot, 2005). It is evident that the parallel processing of input and output information taxes different neurocognitive resources.

During SI, there is maximal use of linguistic and cognitive control hubs compared to simultaneous repetition (Hervais-Adelman et al., 2015). It does not, therefore, come as a surprise that professional interpreters have shown clear advantages in terms of improved memory and EF functions. Professional interpreters seem to exhibit greater WM capacity when compared with comparison groups (i.e., students and non-interpreters) (Mellinger and Hanson, 2019). In a study exploring what professional background can best support respeaking, Szarkowska et al. (2018) suggested that interpreters achieved higher accuracy rating scores in IRSP when compared with translators, and people with no interpreting or translation experience. The difference in IRSP performance was moderated by WM capacity (Szarkowska et al., 2018). In addition, Morales et al. (2015) found that professional SI participants were better at maintaining, updating, and processing of information in the WM when measured with a N-back Task, compared to individuals who were fluent in the second language but had no professional experience. Finally, studies into bilingualism have found that EF skills, including mental flexibility, task switching, and attentional and inhibitory control, are enhanced compared to monolinguals (Soveri et al., 2011; Strobach et al., 2015).

1.3 The effects of interpreter training on WM and performance

Previous studies have shown that interpreter training can boost participants' WM performance (Macnamara and Conway, 2016; Babcock et al., 2017; Chmiel, 2018). Also, nine months of SI training have been shown to cause structural and functional brain changes in temporoparietal, frontostriatal, and temporoparietal brain circuits (Van de Putte et al., 2018). Chmiel (2018) conducted a longitudinal study over 2 years where WM performance (measured with a reading span task) of professional interpreters, interpreter trainees, and bilingual controls, was investigated. The professional interpreters outperformed on WM tasks at baseline. However, after a 2-year interpreter training, the trainees scored higher on WM tasks (Chmiel, 2018). In another longitudinal study, the WM performance (measured with backward span, reading span, and operation span) of 50 American Sign Language (ASL) simultaneous interpreting students was measured before and then 2 years after a sign language interpreting course: the students' simple WM (i.e., backward span) was enhanced, but not their complex WM (i.e., reading span and operation span). Thus, SI training appears not to improve memory skills that require concurrent storing and processing of information (Macnamara and Conway, 2016). Likewise, Babcock et al. (2017) conducted a longitudinal investigation and found that 2 years of SI training was associated with positive cognitive changes in verbal short-term memory, measured with a letter span task. There were no significant findings in relation to operation span and symmetry span (complex measures of WM).

In the SMART project's experiment, a customized training course was created to ensure that all participants, i.e., language professionals from various backgrounds, received equal exposure to IRSP before undergoing testing (see Materials—The IRSP Upskilling Training-For-Testing Course). Given the hybrid nature of IRSP, sharing many similarities with SI, we decided to investigate cognitive enhancement between the start and end of the course. Due to the multi-step nature of IRSP, simple WM was not part of these analyses.

1.4 Positive cognitive changes through specific skills training

Several studies in domains other than Translation and Interpreting Studies have investigated how training in specific skills can lead to positive cognitive changes. One of these critical research areas is online gaming and action games, in particular, where cognitive functions can be enhanced by the extensive practice of playing the games (Boot et al., 2011). Online action games are comparable to IRSP in that both involve multiple simultaneous actions that occur in real time. EF skills in these domains involve processing complex situations involving simultaneous and sequential tasks with quick, real-time switches between them (Logan and Gordon, 2001). Frequent video gaming has been found to benefit the development of EF skills, particularly attention skills, task switching, and WM (Alho et al., 2022) suggesting similarities between the cognitive requirements of online gaming and IRSP. Many studies have shown that when young adult non-gamers are trained in action video games, their visual attention skills, task switching, and multiple object tracking improve (Green and Bavelier, 2003; Strobach et al., 2012; Oei and Patterson, 2013). In another study, Parong et al. (2017) tested a custom-made online game (Alien Game) that focused on EF skills, concluding that playing 2 h of the online game when compared to a control game could improve shifting skills (Parong et al., 2017).

1.5 Study hypotheses

The current study is looking to investigate how cognitive processes of language professionals are affected when working on cognitively demanding real-time multi-step processes that rely on HAII. To this end, we investigated how cognitive resources measured at baseline i.e., reading and digit span (WM), N-back (a measure of maintaining, updating, and processing of information in the WM), switching skills, and sustained attention were associated with high IRSP performance that was measured at the end of the IRSP training course. This was to further extend and substantiate the initial findings by Pöchhacker and Remael (2020), providing empirical evidence to some of the essential skills and competences required in the IRSP process. Additionally, we wanted to further explore how the purpose-built IRSP training course might affect the wider cognitive skills (complex WM, switching skills, and sustained attention) of language professionals. Notably, previous empirical findings on simultaneous interpreters have indicated that complex WM can be enhanced (Chmiel, 2018). Given the multi-step nature of IRSP, we anticipated similar advantageous effects and benefits in this domain as well.

Therefore we hypothesized that there would be a positive relationship between baseline complex WM resources and post-training IRSP accuracy (Hypothesis 1) in line with previous findings on SI (Timarova, 2007; Aben et al., 2012). With respect to N-back (WM) we predicted that it would be positively associated with high IRSP accuracy (Hypothesis 2). Previous findings have suggested that participants with professional SI experience outperform control participants with fluency in the second language but no professional experience of SI, on N-back performance (Morales et al., 2015). Also, given that several cognitive abilities in relation to IRSP have not been tested previously, this study took an exploratory approach to investigate these further. Hence, we investigated how simple WM, switching skills, and sustained attention might predict IRSP accuracy.

Furthermore, the effects of the training course on cognitive performance were examined. It was predicted that after attending a 5-week training course on IRSP, there would be an enhancement on complex WM (Hypothesis 3) as suggested by Chmiel (2018). We also hypothesized (Hypothesis 4) that switching skills would be improved after the training because of evidence from other cognitively similar domains, online gaming in particular (Parong et al., 2017; Alho et al., 2022). Our final hypothesis (Hypothesis 5) was that sustained attention would improve between the start and end of the training course as many studies in bilingualism have highlighted that attentional and inhibitory control can be improved when compared to monolinguals (Soveri et al., 2011; Strobach et al., 2015).

2 Materials and methods

2.1 Participants

Fifty-one language professionals with English, French, Italian, or Spanish as their mother tongue participated in this study (*M*age = 40.12 years, *SD* = 10.97 years). There were eight males (*M*age = 37.38 years, *SD* = 10.93 years) and 43 females in the study (*M*age = 40.63 years, *SD* = 11.51 years). The participants had a minimum of 2,000 h of professional experience in one or more language-related practices: spoken language interpreting (consecutive) 58.82%; spoken language interpreting (simultaneous) 52.94%; written translation 94.12%; pre-recorded subtitling 58.82%, and/or live subtitling 21.57%. The participants were grouped based on their language directionality: French (nine working into English and eight working into French); Italian (16 working into Italian and one working into English); and Spanish (eight working into Spanish and nine working into English).

2.2 Materials

2.2.1 The IRSP upskilling training-for-testing course ("Advanced introduction to interlingual respeaking")

This paper focuses on data collected before and after participants completed a bespoke 25-h upskilling course, delivered online over 5 weeks and in a self-taught manner. The course had the dual purpose of collecting data for the study (hence training-for-testing) and placing all participants on a level playing field in relation to this practice by providing them an "Advanced introduction to IRSP." Due to the innovative nature of the practice and the limited number of fully trained professionals available, the study team designed the course to cater to language professionals from diverse walks of life, each bringing unique skills to this emerging field. To this end, the course broke down interlingual respeaking into three key modules: on technology, particularly exploring the main components of speech recognition software (Dragon Naturally Speaking v 15) and its functioning; intralingual practice, i.e., in the same language; and interlingual practice, i.e., into another language. The course proceeded through four sequential blocks that guided the learners through the steps required for IRSP gradually: (1) Simultaneous listening and speaking/translating and software-adapted delivery (i.e., how to adjust one's voice and prosody to ASR for optimal recognition); (2) Adding punctuation and related strategies for chunking and dealing with speed; (3) Software optimization and preparation prior to a respeaking task for accuracy; (4) Error correction via different methods. Learning proceeded through alternation of theory and practical exercises, designed to train each procedural skill firstly independently then in combination with others, in an incremental way. Participants performed each task first intralingually, then interlingually, before moving on to the next one, which allowed the participants to train in an incremental manner across a predetermined sequential order. Each task had to be completed before participants were permitted to proceed to the next one. At the end of the course, participants were tested on both intralingual and interlingual respeaking.

2.3 Cognitive measures

2.3.1 WM (reading span task)

The reading span task (RST) is a complex memory span task including a processing component (lexical decision: judging the correctness of sentences) and a storage component (memorizing a series of words for subsequent recall) (Daneman and Carpenter,

1980). RST is widely used and adapted for verbal WM and cognitive processing investigations. It focuses on the active updating and monitoring of information in WM. Before the actual RST comprising 12 blocks, there were three practice trials. The task contained between 2 and 5 sentences in each block, and the participants were asked to judge the correctness of the sentences (e.g., "The surgery's giraffe is arriving after 20 min to open the doors" or "The mother rushed to the school to pick up her daughter"). In the storage component, there were between 2–5 words (e.g., "pet" and "bug") to be recalled later. The primary output measure of the RST was the recall proportion of the words remembered (i.e., storage component of the task). The score on the correctness of sentences was not measured. It was used to make sure that the participants were paying attention to the task. The same RST was used during the pretesting and post-testing phases of the experiment. However, the sentences and words used were different during the pretesting and post-testing stages to avoid any practice effects. The participants completed an online version of the RST that was created in Pavlovia.

2.3.2 WM (digit span task)

The digit span task (DST) is a simple memory span task. Unlike the complex WM measure that measures both processing and storage of WM, simple memory span task focuses on WM storage only. In this task, a person is presented with a sequence of digits (starting three digits) and asked to repeat the sequence. Participants do three conditions as part of the DST: forward span where the digits are recalled in the same order, backward span where the participants need to recall the digits in the backward order, and then recalling of digits in an ascending order involving the participant to sequence the numbers from the lowest to the highest. The number of digits increases 1 at a time (two trials for each span) until the participant fails on both trials. The longest remembered sequence is the person's digit span for that condition. This task was also created in Pavlovia and the participants completed it online.

2.3.3 Switching skills

Switching skills were measured with a plus-minus task which measures switching between simple mathematical operands of addition and subtraction (Miyake et al., 2000). This function focuses on shifting back and forth between multiple tasks or mental sets and it can also be called attention switching and task switching. The participants started with addition, moving into subtraction, and finished with a task where they alternated between additions and subtractions. All the numbers used in the task were two-digit numbers (from 10 to 99), and they were only used once. The numbers (30 per condition, presented in a vertical column) were randomly mixed to form the three conditions (i.e., addition, subtraction, and switch: alternation between addition and subtraction). Participants worked their way down the column and entered the answer in the space next to it, in Qualtrics. Time taken was measured, when they completed a column, Qualtrics moved on to the next condition. First, they added the number three to each number (e.g., 83 + 3, addition condition. Then, for the second condition, they subtracted the number three (e.g., 75 − 3). Then, they alternated between addition and subtraction of a 3 as they worked their way down the column, in the third condition. A switching cost was calculated where the non-switch completion time (an average of time taken to complete the addition and subtraction conditions) was subtracted from the time to complete the switch condition). The same plus-minus task was used during the pretesting and post-testing. However, the randomization of the double-digit numbers was different during the pretesting and post-testing stages of the experiment to mitigate any practice effects.

2.3.4 Sustained attention to response task

In this computer-based go/no go task, participants are required to make a response every time they see a number (1–9) by pressing a key, except when that number is three, in which case they must withhold their response (Robertson et al., 1997; Manly and Robertson, 2005). During the sustained attention to response task (SART) task, inhibitory control is necessary to discriminate between relevant and irrelevant distractors (Manly and Robertson, 2005). Sustained attention is required for constant monitoring of the task. Five blocks of 45 trials each (225 trials in total) were presented visually over 4.3 min. The participants responded with a key press to each digit except when the number three appeared on the screen (25 times) when they had to withhold their response. Number three was distributed throughout the 225 trials in a quasi-random way. The participants used their preferred hand to respond and were told to focus on accuracy and speed equally. Before starting the actual task, each participant did a practice that comprised eighteen numbers, two of which were the target number three. The primary measure of the SART task was the proportion of targets ("3") to which the participants successfully withheld their response. We also measured the average reaction time (in seconds) of the participants. This online version of the task was created in Pavlovia.

2.3.5 N-back task

N-back is a widely used measure for assessing WM, which requires the participant to maintain, continuously update, and process information (Kirchner, 1958). N-back is commonly used to measure WM monitoring and updating, while minimizing the storage component (Morales et al., 2015). Hence, it is used to evaluate the updating function of the Miyake's model of EF (Miyake et al., 2000) and is linked to the central executive (CE) of the Baddeley's model of WM where it refers to the monitoring of incoming information for task relevance. The information that is not needed for the completion of the task is updated with the new information as part of the CE (Baddeley and Hitch, 1974; Baddeley, 2012). In the current study, the participants completed two practice blocks before the actual task: one for the 0-back and the second one for the 2-back. Participants were instructed to monitor a series of stimuli and to respond whenever a stimulus was presented that was the same as the one presented *n* trials before. The letters that acted as the stimuli were presented for 500 ms followed by a 2,500 ms black period. The N-back Task had an equal number of blocks for 0-backs (10 blocks) and 2-backs (10 blocks). Participants either matched a letter to the target (0-back) or indicated whether it matched with one presented 2 before (2-back) by pressing a key on the keyboard. Average accuracy was calculated for 0-back blocks and 2-back blocks. This was an online task created in Pavlovia.

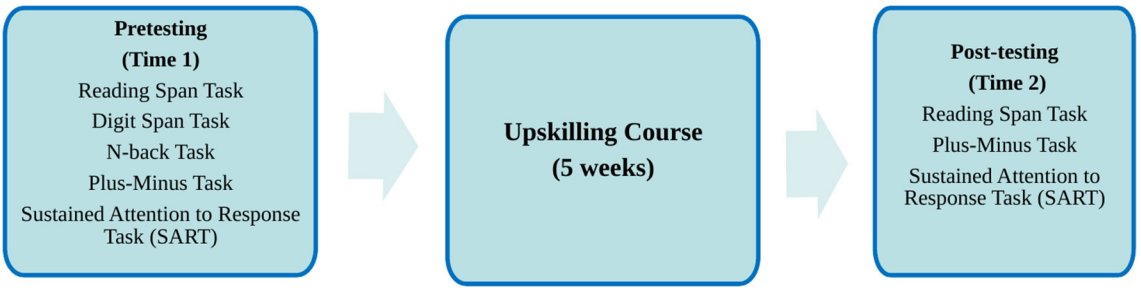

FIGURE 1
The procedure of the experiment (cognitive component).

TABLE 1 Descriptive statistics for all the variables (pretesting and post-testing data).

Measures	Pretesting			Post-testing		
	M	*SE*	*P*-value	*M*	*SE* (SD)	*P*-value
RST	0.83	0.02	0.05	0.88	0.02	0.05
Plus-minus (s)	22.90	2.95	0.02	14.55	1.85	0.02
SART (accuracy)	0.96	0.003	0.50	0.96	0.004	0.50
SART (RT in s)	0.37	0.008	0.06	0.39	0.009	0.06
NTR accuracy (%)				95.37	(1.5)	

RST, reading span task (WM); Plus-minus, plus-minus task (switching skills measured with a switching cost in seconds); SART, sustained attention to response task (sustained attention). NTR accuracy was only measured after the training course.

TABLE 2 Multiple regression model with IRSP accuracy as the criterion variable.

Measures	Unstandardized Coefficients (B)	Standardized Coefficients (beta)	*t*	*P*-value
RST	0.03	0.32	2.33	0.03
DST (1)	0.00	−0.04	−0.25	0.80
DST (2)	0.00	0.23	1.43	0.16
DST (3)	−0.00	−0.11	−0.74	0.47
N-back	0.02	0.20	1.46	0.15
Plus-Minus (s)	0.00	−0.14	−1.04	0.30
SART (accuracy)	−0.14	−0.21	−1.54	0.13

RST, reading span task (WM); DST, digit span task (WM); DST 1, forward span; DST 2, backward span; DST 3, ascending numerical order; N-back, N-back task (WM); Plus-minus, plus-minus task (switching skills measured with a switching cost in seconds); SART, sustained attention to response task (sustained attention).

2.4 Procedure

Due to the pandemic, this study was entirely conducted online and advertised on the SMART project homepage and dedicated social media account. Any interested language professionals were sent an eligibility questionnaire that focused on language eligibility (i.e., English paired with Italian, Spanish, and/or French, with at least one of these languages as their mother tongue), professional eligibility (i.e., a minimum of 2,000 h of professional experience in language-related practices, namely consecutive and/or dialogue interpreting, translation, live and/or offline subtitling), and correct equipment specifications (laptops, headset, and microphone). Participants who met all the eligibility criteria were sent a link to the study, comprising cognitive tasks created in Pavlovia and Qualtrics. Consent was given by the participants before starting. The experiment started as soon as the participant opened the link, and it took the participant through the whole pretesting phase of the experiment in one go (Figure 1 summarizes the procedure of the experiment in relation to the cognitive component analyzed in this paper). Before beginning the data collection process, we integrated Pavlovia and Qualtrics and tested it carefully to ensure that all participants would go through the same experimental steps. Despite the lack of strict experimental conditions, we aimed for a rigorous approach. We also had a pre-testing call with each participant to explain the cognitive testing steps and answered any questions. Participants were pretested on several cognitive

abilities, specifically WM (including reading span, digit span, and N-back), switching skills, and sustained attention, with a duration of 40 min. The pretesting was followed by the 25-h upskilling course. The participants were provided with a link to the upskilling course, which was hosted on Moodle, and worked on the four different blocks independently online. Subsequently, they were tested on their intra and interlingual respeaking performance. In the current study, only the interlingual respeaking performance was used as a basis of accuracy for our investigations. The accuracy of the subtitles thus produced was assessed using the NTR model (Romero-Fresco and Pöchhacker, 2017—see Analytic plan below), which focuses on the type of errors made while performing IRSP. After the training-for-testing, participants were sent a follow-up link to complete three post-testing cognitive measures (reading span, switching skills, and sustained attention), which took ∼25 min and were delivered via the same platform as the one used for pre-testing measures (Pavlovia). Upon completion of the cognitive tasks, participants were asked to take part in a final evaluation questionnaire to gather information and feedback about the course, after which they were debriefed and thanked for their participation.

2.5 Analytic plan

Our study involved a two-part statistical analysis that enabled us to examine our five hypotheses. In the first phase of the analysis, we concentrated on IRSP accuracy measured at the end of the training and how it was predicted by baseline cognitive abilities.

To assess the accuracy of IRSP performance, we employed the NTR Model (Romero-Fresco and Pöchhacker, 2017) which specifically focuses on the nature of errors committed by language professionals while producing interlingual live subtitles via respeaking. The NTR formula distinguishes between software-related recognition and human translation errors, including content-related errors (i.e., omissions, additions, and substitutions) and form-related errors (grammatical correctness and style). NTR accuracy is based on the following formula: NTR = $[(N\text{-}T\text{-}R)/N] \times 100\%$, where N, number of words; T, translation errors; and R, recognition errors. Errors get different scores depending on their severity. Minor errors are penalized with a −0.25-point deduction as they do not impair comprehension. Major errors, however, can cause confusion and information loss, and are penalized with a −0.50-point deduction. Finally, critical errors can introduce false or misleading information, and therefore they are penalized with a −1.0-point deduction. Intralingual subtitles (i.e., in the same language) are required to reach an accuracy rate of 98%. A similar accuracy rate is suggested for interlingual subtitles (i.e., in a different language), although this benchmark has not been validated yet.

Multiple regression was used to investigate what predicted post-training IRSP performance. Our multiple regression model predictors were reading span (WM), digit span (WM), N-back (WM), switching skills, and sustained attention, at baseline.

In the second (longitudinal) part of the analysis, we looked at the effects of the IRSP course on three cognitive abilities that were measured both before and after the course (reading span, switching skills, and sustained attention) using a repeated-measures within-subjects design by looking at changes in cognitive performance between these two time points.

3 Results

Before conducting the actual statistical analyses, we investigated whether the data was normally distributed. Shapiro-Wilk's test was non-significant for all the variables, suggesting that all the data were normally distributed. We also viewed histograms to confirm normality and checked box plots. No extreme values or outliers were found. The test of sphericity was non-significant ($p > 0.05$) indicating that the assumption of sphericity was met. Table 1 summarizes all the descriptive statistics (pretesting and post-testing data). Based on the NTR Model, the participants' average IRSP accuracy was $M = 95.37\%$ and $SD = 1.5\%$, indicating that the accuracy was lower than the recommended 98% for intralingual live subtitles (i.e., in the same language). Table 2 includes all the data for the multiple regression analysis.

3.1 Cognitive predictors of IRSP accuracy

A multiple linear regression was conducted to predict NTR accuracy based on pre-testing (baseline) reading span, digit span, N-back, switching skills, and sustained attention. The multiple regression model was significant $F_{(7,42)} = 2.27$, $p = 0.04$ and the adjusted R^2 indicated that 15.4% of the variance in the IRSP accuracy was explained by the model. There was a significant positive relationship ($\beta = 0.32$) between the participants' reading span (a complex WM measure) and their IRSP accuracy. However, the other predictors (i.e., digit span, N-Back, switching skills, and sustained attention) were not statistically significant, as $p > 0.05$ (see Table 2).

3.2 Pretesting and post-testing data

To investigate the possible effects of the IRSP training course on cognitive performance, we compared the cognitive performance of participants from before (pretesting—baseline, T1) to after the course (post-testing, T2). Our variables were reading span (a complex WM measure), switching skills, and sustained attention.

3.2.1 Complex WM (reading span)

Complex WM from the RST task improved from T1 to T2, $F_{(1,46)} = 4.0$, $p = 0.05$ (from $M = 0.83$, $SE = 0.02$ to $M = 0.88$, $SE = 0.02$), suggesting that the IRSP training course might have improved the WM resources of the language professional (Figure 2).

3.2.2 Switching skills

Switching skills (switching cost in seconds from the plus-minus task) improved from T1 to T2, $F_{(1,49)} = 6.42$, $p = 0.02$ (from $M = 22.90$ s, $SE = 2.95$ s to $M = 14.55$ s, $SE = 1.85$ s) showing that

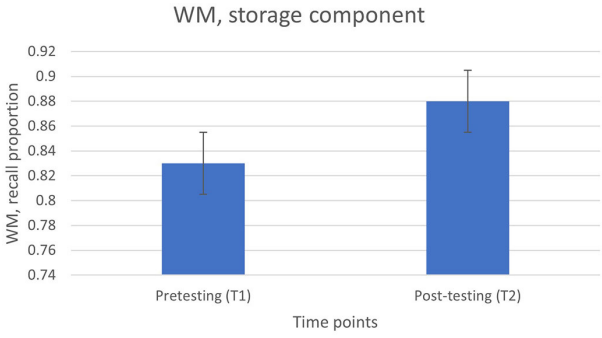

FIGURE 2
WM at time points 1 and 2.

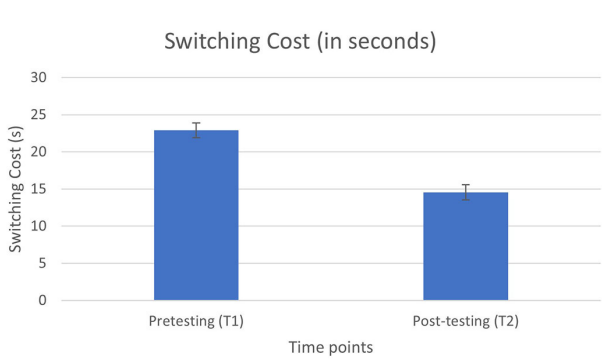

FIGURE 3
Switching cost at time point 1 and at time point 2.

the IRSP training course might have enhanced the participants' switching skills (Figure 3).

3.2.3 Sustained attention

There was no significant change in sustained attention (SART), accuracy between T1 and T2, $p > 0.05$. However, there was a marginally significant difference in SART reaction times (seconds) between T1 ($M = 0.37$, $SE = 0.008$) and T2 ($M = 0.39$, $SE = 0.009$), $F_{(1,48)} = 3.60$, $p = 0.06$, with the means indicating that the participants were slower on the SART task post-training.

4 Discussion

AI-related technologies are developed with the goal of augmenting and improving human efficiency. However, at present, human involvement is still necessary to oversee and modify the output generated by these technologies. As a result, this places an additional burden on humans, increasing the number of tasks they are responsible for managing. When working with AI-related technologies such as ASR, a new working environment is created in the form of HAII where human cognition is partly distributed to the outside world. IRSP is a case in point, where very little is known about its process and the human cognitive requirements for this newly-created HAII, and in turn how cognitive processes are affected by engaging with it. To this end, we focused on IRSP, a cognitively demanding process, to study the links between this novel form of HAII and human cognition. We investigated what baseline cognitive abilities predicted higher IRSP performance after a 25-h upskilling course. We also explored whether the course would improve the EF and complex WM of language professionals. We focused on these cognitive areas, as previous work on SI and bilingualism had suggested their involvement and highlighted the possibility of improvements within these domains.

Our hypotheses were partly supported. There was a positive relationship between complex WM resources (reading span) measured at baseline and subsequent post-training IRSP performance, confirming our first hypothesis. Complex WM was the only significant finding in relation to the five cognitive predictors of high IRSP performance under investigation (complex WM, simple WM, N-back, switching skills, and sustained attention), clearly emphasizing complex WM as a leading competence required for accurate IRSP performance. These results are in line with existing findings suggesting that WM resources are required to manage the cognitively demanding process of SI, a practice that shares many similarities with IRSP (Timarova, 2007; Aben et al., 2012; Mellinger and Hanson, 2019). Furthermore, this finding complements the process and competence model by Pöchhacker and Remael (2020) by bringing empirical evidence regarding complex WM resources and their role in achieving high accuracy in IRSP. On the other hand, simple working memory, measured with a digit span task (focusing on WM storage only), was not a predictor, suggesting that to perform well in IRSP, simultaneous processing and storage of WM are both required rather than just WM capacity *per se*. The multi-step and real-time nature of IRSP is likely the reason why both processing and storage are required, to enable professionals to keep up with the task and reach high accuracy.

There were no significant findings in relation to the third WM measure (N-back Task) failing to support our second hypothesis. These findings are in contrast with Morales et al. (2015) who suggested that participants with SI experience performed better in monitoring and updating, measured with the N-back Task, when compared to a control group. It is possible that the use of ASR as part of the IRSP process alters the WM requirements, leading to a stronger need for simultaneous processing and storing of information (measured with a complex WM task) rather than just updating of information. The monitoring and possible editing of the ASR output might change the focus of the language professional and therefore create an environment where strong complex WM resources are more important. Also, it should be noted that both the N-back and digit span tasks use numeric rather than word stimuli. This could have contributed to the lack of effects and future studies should consider adapting the tasks to explore whether this is important in this context.

Additionally, our findings highlighted that switching skills were not positively associated with better post-training IRSP performance. This finding is somewhat surprising as past findings have indicated that bilinguals have enhanced switching skills compared to monolinguals (Soveri et al., 2011) and that SI improves the ability to coordinate multiple tasks in dual-task situations (Strobach et al., 2015). It might be that the plus-minus task used here (as a pure "switching" measure) does not

capture the complexity and variety of the multitasking skills required by IRSP, and a longer and more advanced task requiring switching between multiple sources of information would have found effects. However, time constraints precluded use of such task here.

Similarly, sustained attention was not associated with high IRSP accuracy either. IRSP is a time-pressured process with high levels of task demand. Its continuous demands on cognitive resources meant that we expected sustained attention to be a predictor, but this was not supported. However, continuous performance tasks such as SART require subjects to maintain attention during a monotonous, repetitive, task with low levels of demand. Again, this does not reflect the IRSP environment. It seems that the ability to avoid distraction and mind-wandering during such a task is not a predictor of IRSP performance, but this is perhaps not surprising when IRSP imposes such high cognitive demands. Also, it should be noted (as discussed further below) that SART performance was very high across the sample, so ceiling effects could explain the lack of predictive utility. Continuous attention when task demands are high might be a more appropriate measure which should be tested in future work. In sum, results highlight the importance of complex WM as a predictor of IRSP accuracy, with simple WM, switching skills, N-back (maintaining, updating, and processing of information in WM), and sustained attention not being significant predictors. Future studies should explore the role of complex WM in more detail and include alternate measures of the other cognitive skills to confirm the current findings.

When looking at the effects of the IRSP training course on cognitive performance, our results suggested that complex WM improved between the start and end of the training course, indicating that there can be possible cognitive enhancements with IRSP training, confirming our third hypothesis. These findings also highlight the malleability of WM resources with the help of a training course, confirming existing findings around effects of cognitive training (Morrison and Chein, 2011; Pappa et al., 2020). As aforementioned, the use of ASR might change the cognitive environment with more emphasis on the requirement of complex WM resources. By attending the training course, complex WM resources were challenged and seem to have improved. These findings are in line with Chmiel (2018) who confirmed that after a 2-year training in interpreting, the interpreter trainees scored higher than professional interpreters on complex WM tasks. However, these results do not align with previous investigations by Macnamara and Conway (2016) who reported that a 2-year SI training in ASL did not improve complex WM of ASL simultaneous interpreting students, but the training did enhance simple WM resources. Similarly, there were no significant findings in relation to complex WM measures but to simple WM measures after 2 years of SI training in the Babcock et al. (2017) study where the performance of SI students was compared to translation students and non-language students as the control groups. It is possible that the multi-step nature of IRSP, including the use of ASR might explain the improvement of complex WM in our study in contrast to Babcock et al. (2017) and Macnamara and Conway (2016) studies. Any future studies should focus on looking at the complex WM resources of different groups of language professionals.

In the present study, although switching did not emerge as a predictor of post-training IRSP performance, switching skills were enhanced after the IRSP training course, confirming our fourth hypothesis. IRSP requires the language professional to actively switch between tasks, involving simultaneous interpreting, and monitoring of the ASR output. The design adopted in the IRSP training course has, therefore, facilitated the development of these skills among language professionals, possibly leading to their enhancement. Similar to complex WM, our findings support the idea that switching skills are malleable (Zhao et al., 2020) and can be enhanced with training. However, the shortcomings of the plus-minus task and the fact that it did not predict performance, means this result should be viewed with caution: it could be an artifact of task-specific practice effects. Nevertheless, these findings align with current research findings in online gaming as multi-step process activities (Parong et al., 2020; Alho et al., 2022). Frequent online gaming was found to benefit task switching (Alho et al., 2022) and shifting between competing tasks (Parong et al., 2017). Finally, our findings have clearly highlighted that new forms of HAII might increase the number of tasks the human needs to focus on; however, these findings also indicate potential cognitive benefits for the individual engaging in this complex practice.

Regarding our final hypothesis, which posited improvements in sustained attention accuracy, there were no significant findings, thus failing to support the hypothesis. The baseline SART accuracy was high (96.1%), possibly indicating that the SART task was easy for the language professional to complete because of possible previous experience in activities requiring sustained attention. Perhaps, there were ceiling effects and therefore, the accuracy could not be enhanced any further. However, when looking at the SART reaction times, the language professionals became slower (at trend) post-training. IRSP process fosters a behavior where the accuracy of the subtitles produced is imperative. This has perhaps led the language professional to be more prudent with their strategies while completing the SART task, leading to slower reaction times (RT). Similar pattern was found by Vallesi et al. (2021) who suggested that SART accuracy was improved with additional vigilance with the task. It is also possible that another type of attention is required in IRSP and that is why future research should focus on other types of attentional skills (e.g., divided attention).

In the present study, DCog is seen as the foundation of HAII. Clearly, a new cognitive environment is created with IRSP where parts of the human cognition is distributed with the help of ASR, leading to interactions between humans and technological tools. According to DCog, the use of external artifacts and technology have the potential to increase the workspace available for the human (Kirsh, 2010; Vallée-Tourangeau, 2013; Wallinheimo et al., 2019). However, it is not clear what happens during the IRSP process when technology does not work the way the language professional wants it to. During the IRSP process, the language professional might need to correct what has been produced by the ASR and it is possible that the human loses the sense of personal control over the situation (Ehrensberger-Dow and O'Brien, 2015). There can be additional worry and anxiety, leading to additional taxation of WM and hampered IRSP performance. Any future experimental IRSP studies should focus on these important aspects that allow humans and technology to work successfully together.

4.1 Limitations and future studies

Whilst we have revealed some interesting findings that advance literature, there are clear limitations. In IRSP, there are additional steps for the language professional at the core of the activity to monitor and ensure the accuracy of the subtitles produced in conjunction with ASR technology. We suspect that this might lead to an increased workload and cognitive load. However, we have not measured cognitive load as part of the present investigation. Future studies should focus on understanding how the different tasks requiring varying cognitive resources affect the human's cognitive load and whether this impairs respeaking performance and other cognitive performance. This approach could then be transferred to other real-time HAII practices witnessing high burnout risks (e.g., the financial sector and aviation industry), allowing for optimal performance without ignoring the needs of the individual involved.

From a methodological perspective, it is noteworthy that the entire study was carried out online due to the pandemic. Despite our efforts to create a seamless and well-integrated experience for participants, as detailed in the procedure section, variations in participants' individual testing environments during the experiment are possible. However, we took measures to minimize potential repercussions on the conduct of the experiment. We ensured that all tasks were organized within a clear and structured flow, complete with instructions. Moreover, we communicated directly with participants before the tests (via individual pre-testing calls), emphasizing the importance of completing them in a quiet environment without disruptions to avoid breaking the flow and getting distracted. We closely monitored the process by focusing on the reaction times and found no indications of participants not adhering to the provided instructions.

In addition, it is possible that there were practice effects on the cognitive tasks between the pre-testing and post-testing phases when reading span, switching skills, and sustained attention were measured. Our investigation is focused on language professionals who completed an upskilling course on IRSP. However, we have not compared our findings to a control group. Future research should focus on investigating any possible cognitive changes in relation to other similar types of training courses compared to the IRSP training of language professionals. Our study involved a professional sample with an older average age. Therefore, comparing our findings with other studies that mainly focus on student samples might be challenging. However, it is true to say that an older sample might be more motivated to participate in a study like this (Ryan and Campbell, 2021). Additionally, while we used cognitive measures that have been previously used in SI research, it is important to note that SI rely on a different degree of interaction with technology, and thus creates a different cognitive environment when compared to IRSP. As such, different cognitive measurements might be needed to effectively evaluate human-AI collaboration in this practice.

5 Conclusion

The present study has allowed us to complement and provide empirical evidence to the process and competence models by Pöchhacker and Remael (2020) by suggesting that complex WM resources are required to achieve high IRSP accuracy. These findings could be transferred to other similar real-time work processes involving humans and technology to highlight the importance of complex WM resources in comparable practices. Furthermore, our study adds to the growing literature on possible cognitive enhancements after a training course. We found that both complex WM and switching skills were improved with IRSP training, highlighting the fact that these skills can be trained and their possible malleability. The newly-created HAII environment of IRSP seems to lead to positive cognitive enhancements for the language professional. Whilst there might be an increased workload by monitoring and editing the output of ASR during IRSP, there seem to be clear cognitive benefits in doing so. However, more investigations are required to further understand the possible risk of burnout when working in real-time HAII practices to allow for AI that is fully people-centered and responsible. This approach would also support the International Labor Organisation's (ILO) Decent Work agenda that helps advance all employees' working conditions in varied working environments.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by University of Surrey (UK) Ethics Committee. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

A-SW: Data curation, Formal analysis, Writing – original draft. SE: Conceptualization, Methodology, Writing – review & editing. ED: Conceptualization, Methodology, Resources, Writing – review & editing.

Funding

This study was part of the ESRC-funded SMART project (Shaping Multilingual Access through Respeaking Technology, ES/T002530/1, 2020–2023).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Aben, B., Stapert, S., and Blokland, A. (2012). About the distinction between working memory and short-term memory. *Front. Psychol.* 3, 301. doi: 10.3389/fpsyg.2012.00301

Alho, K., Moisala, M., and Salmela-Aro, K. (2022). Effects of media multitasking and video gaming on cognitive functions and their neural bases in adolescents and young adults. *Eur. Psychol.* 27, 131–140. doi: 10.1027/1016-9040/a000477

Babcock, L., Capizzi, M., Arbula, S., and Vallesi, A. (2017). Short-term memory improvement after simultaneous interpretation training. *J. Cogn. Enhanc.* 1, 254–267. doi: 10.1007/s41465-017-0011-x

Babcock, L., and Vallesi, A. (2017). Are simultaneous interpreters expert bilinguals, unique bilinguals, or both? *Biling.: Lang. Cogn.* 20, 403–417. doi: 10.1017/S1366728915000735

Baddeley, A. (2012). Working memory: theories, models, and controversies. *Annu. Rev. Psychol.* 63, 1–29. doi: 10.1146/annurev-psych-120710-100422

Baddeley, A. D., and Hitch, G. (1974). "Working memory," in *The Psychology of Motivation VIII*, ed G. Bower (New York, NY: Academic Press), 47–89. doi: 10.1016/S0079-7421(08)60452-1

Boot, W. R., Blakely, D. P., and Simons, D. J. (2011). Do action video games improve perception and cognition? *Front. Psychol.* 2, 262. doi: 10.3389/fpsyg.2011.00226

Chmiel, A. (2018). In search of the working memory advantage in conference interpreting – training, experience and task effects. *Int. J. Biling.* 22, 371–384. doi: 10.1177/1367006916681082

Christoffels, I., and de Groot, A. (2005). "Simultaneous interpreting: a cognitive perspective," in *Handbook of Bilingualism: Psycholinguistic Approaches*, eds J. Kroll, and A. de Groot (Oxford: Oxford University Press), 454–479.

CSA Research (2023). *Forecast 2023: Language Market*. CSA Insights. Available online at: https://insights.csa-research.com/reportaction/305013489/Marketing (accessed June 2023).

Daneman, M., and Carpenter, P. A. (1980). Individual differences in working memory and reading. *J. Verbal Learn. Verbal Behav.* 19, 450–466. doi: 10.1016/S0022-5371(80)90312-6

Davitti, E., and Sandrelli, A. (2020). Embracing the complexity. A pilot study on interlingual respeaking. *J. Audiov. Transl.* 3, 103–139. doi: 10.47476/jat.v3i2.2020.135

Ehrensberger-Dow, M., and O'Brien, S. (2015). Ergonomics of the translation workplace. *Transl. Spaces* 4, 98–118. doi: 10.1075/ts.4.1.05ehr

Green, S., and Bavelier, D. (2003). *Action Video Game Modifies Visual Selective Attention*. Available online at: https://www.nature.com/nature (accessed July 2023).

Hervais-Adelman, A., Moser-Mercer, B., Michel, C. M., and Golestani, N. (2015). fMRI of simultaneous interpretation reveals the neural basis of extreme language control. *Cereb. Cortex* 25, 4727–4739. doi: 10.1093/cercor/bhu158

Hollan, J., Hutchins, E., and Kirsh, D. (2000). Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Trans. Comput. Hum. Interact.* 7, 174–196. doi: 10.1145/353485.353487

Hutchins, E. (1995). *Cognition in the Wild*. Cambridge, MA: MIT Press, 175–228. doi: 10.7551/mitpress/1881.001.0001

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *J. Exp. Psychol.* 55, 352–358. doi: 10.1037/h0043688

Kirsh, D. (1995). The intelligent use of space. *Artif. Intell.* 73, 31–68. doi: 10.1016/0004-3702(94)00017-U

Kirsh, D. (2010). Thinking with external representations. *AI Soc.* 25, 441–454. doi: 10.1007/s00146-010-0272-8

Logan, G. D., and Gordon, R. D. (2001). Executive control of visual attention in dual-task situations. *Psychol. Rev.* 108, 393–434. doi: 10.1037/0033-295X.108.2.393

Macnamara, B. N., and Conway, A. R. A. (2016). Working memory capacity as a predictor of simultaneous language interpreting performance. *J. Appl. Res. Mem. Cogn.* 5, 434–444. doi: 10.1016/j.jarmac.2015.12.001

Manly, T., and Robertson, I. H. (2005). The sustained attention to response test (SART). *Neurobiol. Attention* 337–338. doi: 10.1016/B978-012375731-9/50059-8

Mellinger, C. D., and Hanson, T. A. (2019). Meta-analyses of simultaneous interpreting and working memory. *Interpreting* 21, 165–195. doi: 10.1075/intp.00026.mel

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., and Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: a latent variable analysis. *Cogn. Psychol.* 41, 49–100. doi: 10.1006/cogp.1999.0734

Morales, J., Padilla, F., Gómez-Ariza, C. J., and Bajo, M. T. (2015). Simultaneous interpretation selectively influences working memory and attentional networks. *Acta Psychol.* 155, 82–91. doi: 10.1016/j.actpsy.2014.12.004

Morrison, A. B., and Chein, J. M. (2011). Does working memory training work? The promise and challenges of enhancing cognition by training working memory. *Psychon. Bull. Rev.* 18, 46–60. doi: 10.3758/s13423-010-0034-0

Oei, A. C., and Patterson, M. D. (2013). Enhancing cognition with video games: a multiple game training study. *PLoS ONE* 8, e58546. doi: 10.1371/journal.pone.0058546

O'Neill, T., McNeese, N., Barron, A., and Schelble, B. (2022). Human–autonomy teaming: a review and analysis of the empirical literature. *Hum. Factors* 64, 904–938. doi: 10.1177/0018720820960865

Pappa, K., Biswas, V., Flegal, K. E., Evans, J. J., and Baylan, S. (2020). Working memory updating training promotes plasticity andamp; behavioural gains: a systematic review andamp; meta-analysis. *Neurosci. Biobehav. Rev.* 118, 209–235. doi: 10.1016/j.neubiorev.2020.07.027

Parong, J., Mayer, R. E., Fiorella, L., MacNamara, A., Homer, B. D., and Plass, J. L. (2017). Learning executive function skills by playing focused video games. *Contemp. Educ. Psychol.* 51, 141–151. doi: 10.1016/j.cedpsych.2017.07.002

Parong, J., Wells, A., and Mayer, R. E. (2020). Replicated evidence towards a cognitive theory of game-based training. *J. Educ. Psychol.* 112, 922–937. doi: 10.1037/edu0000413

Pöchhacker, F., and Remael, A. (2020). New efforts? A competence-oriented task analysis of interlingual live subtitling. *Linguist. Antverp. New Ser.* 18, 130–143. doi: 10.52034/lanstts.v18i0.515

Risko, E. F., and Gilbert, S. J. (2016). Cognitive offloading. *Trends Cogn. Sci.* 20, 676–688. doi: 10.1016/j.tics.2016.07.002

Robertson, I. H., Manly, T., Andrade, J., Baddeley, B. T., and Yiend, J. (1997). 'Oops!': performance correlates of everyday attentional failures in traumatic brain injured and normal subjects. *Neuropsychologia* 35, 747–758. doi: 10.1016/S0028-3932(97)00015-8

Romero-Fresco, P. (2011). *Subtitling through Speech Recognition*. London: Routledge.

Romero-Fresco, P., and Pöchhacker, F. (2017). Quality assessment in interlingual live subtitling: the NTR model. *Linguist. Antverp.* 16, 149–167. doi: 10.52034/lanstts.v16i0.438

Ryan, A. D., and Campbell, K. L. (2021). The ironic effect of older adults' increased task motivation: Implications for neurocognitive aging. *Psychon. Bull. Rev.* 28, 1743–1754. doi: 10.3758/s13423-021-01963-4

Soveri, A., Rodriguez-Fornells, A., and Laine, M. (2011). Is there a relationship between language switching and executive functions in bilingualism? Introducing a within group analysis approach. *Front. Psychol.* 2, 183. doi: 10.3389/fpsyg.2011.00183

Strobach, T., Becker, M., Schubert, T., and Kühn, S. (2015). Better dual-task processing in simultaneous interpreters. *Front. Psychol.* 6, 1–9. doi: 10.3389/fpsyg.2015.01590

Strobach, T., Frensch, P. A., and Schubert, T. (2012). Video game practice optimizes executive control skills in dual-task and task switching situations. *Acta Psychol.* 140, 13–24. doi: 10.1016/j.actpsy.2012.02.001

Szarkowska, A., Krejtz, K., Dutka, Ł., and Pilipczuk, O. (2018). Are interpreters better respeakers? *Interpret. Transl. Train.* 12, 207–226. doi: 10.1080/1750399X.2018.1465679

Timarova, S. (2007). "Working memory and simultaneous interpreting," in *Translation and Its Others. Selected Papers of the CETRA Research Seminar in Translation Studies* 2007, ed. P. Boulogne (Belgium), 1–28.

Vallée-Tourangeau, F. (2013). Interactivity, efficiency, and individual differences in mental arithmetic. *Exp. Psychol.* 60, 302–311. doi: 10.1027/1618-3169/a000200

Vallée-Tourangeau, G., and Vallée-Tourangeau, F. (2017). "Cognition beyond the classical information processing model: cognitive interactivity and the systemic thinking model (SysTM)," in *Cognition Beyond the Brain*, eds S. J. Cowley, and F. Vallée-Tourangeau (New York, NY: Springer International Publishing), 133–154. doi: 10.1007/978-3-319-49115-8_7

Vallesi, A., Tronelli, V., Lomi, F., and Pezzetta, R. (2021). Age differences in sustained attention tasks: a meta-analysis. *Psychon. Bull. Rev.* 28, 1755–1775. doi: 10.3758/s13423-021-01908-x

Van de Putte, E., De Baene, W., García-Pentón, L., Woumans, E., Dijkgraaf, A., and Duyck, W. (2018). Anatomical and functional changes in the brain after simultaneous interpreting training: a longitudinal study. *Cortex* 99, 243–257. doi: 10.1016/j.cortex.2017.11.024

Wallinheimo, A., Banks, A., and Tenenbaum, H. (2019). "Achievement goals and mental arithmetic: the role of distributed cognition," in *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, eds A. K. Goel, C. M. Seifert, and C. Freksa (Montreal, QB: Cognitive Science Society), 3057–3063.

Webb, S., and Vallée-Tourangeau, F. (2009). "Interactive word production in dyslexic children," in *The 31st Annual Conference of the Cognitive Science Society*, eds. N. A. Taatgen and H. van Rijn (Austin, TX: Cognitive Science Society), 1436–1441.

Zhao, X., Wang, H., and Maes, J. H. R. (2020). Training and transfer effects of extensive task-switching training in students. *Psychol. Res.* 84, 389–403. doi: 10.1007/s00426-018-1059-7

frontiers | Frontiers in Artificial Intelligence

TYPE Review
PUBLISHED 15 January 2024
DOI 10.3389/frai.2023.1272823

OPEN ACCESS

EDITED BY
Uta Wilkens,
Ruhr-University Bochum, Germany

REVIEWED BY
Stephan Kaiser,
Munich University of the Federal Armed Forces, Germany
Alexandre Ardichvili,
University of Minnesota Twin Cities, United States

*CORRESPONDENCE
Gabor Molnar
✉ gabor.molnar@colorado.edu

RECEIVED 04 August 2023
ACCEPTED 16 November 2023
PUBLISHED 15 January 2024

CITATION
Fenwick A, Molnar G and Frangos P (2024) Revisiting the role of HR in the age of AI: bringing humans and machines closer together in the workplace. *Front. Artif. Intell.* 6:1272823. doi: 10.3389/frai.2023.1272823

Revisiting the role of HR in the age of AI: bringing humans and machines closer together in the workplace

Ali Fenwick[1], Gabor Molnar[2]* and Piper Frangos[3]

[1]Hult International Business School, Dubai, United Arab Emirates, [2]The ATLAS Institute, University of Colorado, Boulder, CO, United States, [3]Hult International Business School, Ashridge, United Kingdom

The functions of human resource management (HRM) have changed radically in the past 20 years due to market and technological forces, becoming more cross-functional and data-driven. In the age of AI, the role of HRM professionals in organizations continues to evolve. Artificial intelligence (AI) is transforming many HRM functions and practices throughout organizations creating system and process efficiencies, performing advanced data analysis, and contributing to the value creation process of the organization. A growing body of evidence highlights the benefits AI brings to the field of HRM. Despite the increased interest in AI-HRM scholarship, focus on human-AI interaction at work and AI-based technologies for HRM is limited and fragmented. Moreover, the lack of human considerations in HRM tech design and deployment can hamper AI digital transformation efforts. This paper provides a contemporary and forward-looking perspective to the strategic and human-centric role HRM plays within organizations as AI becomes more integrated in the workplace. Spanning three distinct phases of AI-HRM integration (technocratic, integrated, and fully-embedded), it examines the technical, human, and ethical challenges at each phase and provides suggestions on how to overcome them using a human-centric approach. Our paper highlights the importance of the evolving role of HRM in the AI-driven organization and provides a roadmap on how to bring humans and machines closer together in the workplace.

KEYWORDS

HRM, AI, AI-HRM, humanizing AI, human-AI integration, workplace, digital transformation

1 Introduction

The intersection of human resource management (HRM) and technology has always been a dynamic space, constantly adapting to market forces and technological innovations. Over the past two decades, the field of HRM has undergone radical transformations, embracing cross-functionality and data-driven approaches (e.g., Bresciani et al., 2021; Zhang et al., 2021). However, the emergence of Artificial Intelligence (AI) has brought about a paradigm shift in HRM, further altering the role of HRM professionals in organizations. With their capacity for enhancing system efficiency, advanced data analysis, and innovation opportunities, AI technologies have begun to permeate multiple facets of organizational functioning, including human resource management (Guenole and Feinzig, 2018; Rathi, 2018).

Despite the growing interest of AI in both business and HRM scholarship, there is limited understanding on these challenges and the opportunities of AI to improve HRM functions and provide positive outcomes for the wider organization (e.g., Agrawal et al., 2017; Castellacci and Viñas-Bardolet, 2019). Moreover, limited knowledge exists on human-AI interaction at work and how HRM can bring humans and machines closer together (e.g., Arslan et al., 2022). The lack of human considerations in HRM tech design and deployment can hamper AI digital transformation efforts and prevent humans from trusting AI-driven processes and tools (e.g., De Stefano and Wouters, 2022). Our paper addresses this gap in the literature by providing a contemporary and forward-looking perspective to the strategic and human-centric role HRM plays within organizations as AI becomes more integrated in the workplace.

In light of these dynamics, this paper explores the challenges and opportunities presented by AI in HRM. Our primary focus is on the interplay between technology and humanity, and the critical role HRM plays in aligning these forces as AI continues to be integrated in the organization. Using a human-centric approach, our framework provides suggestions on how to overcome existing challenges specifically in people management, culture, and compliance. We provide practical suggestions for addressing existing and future challenges in AI adoption and usage within the field of HRM.

1.1 Definitions

HRM is increasingly playing a crucial role in the value creation process of organizations (e.g., DiClaudio, 2019). In this paper we use the definition of HRM by Boselie et al. (2021, p. 484) *"HRM involves management decisions related to policies and practices that together shape the employment relationship and are aimed at achieving certain goals."* HRM goals can be bundled (Beer et al., 2015) to achieve certain organizational outcomes (such as organizational effectiveness and financial performance) or employee/societal centric outcomes (such as well-being). HRM is often operationalized as a combination of different HRM practices together shaping the various employee relationships that exist in and around the organization (Boselie et al., 2021).

Before addressing the changing role of HRM, we first must define AI because without a clear understanding of the term, it is challenging to discern how HRM practices can effectively harness its potential. Existing definitions (e.g., Afiouni, 2019; Lee et al., 2019; Schmidt et al., 2020; Mikalef and Gupta, 2021) generally converge into two main descriptions: (i) the ability to think, understand, and problem-solve like a human, and (ii) the ability to mimic human thinking. It is also important to clarify the terms "artificial" and "intelligence" when defining AI. "Artificial" typically encompasses anything created by humans (e.g., Simon, 1996; Mikalef and Gupta, 2021). On the other hand, "intelligence" refers to a computer's capability to learn, understand, and reason independently, similar to a human (Russell and Norvig, 2010). Nevertheless, there is currently no widely accepted consensus on precisely defining intelligence (e.g., Wang, 2019). Instead, more philosophical notions of intelligence, such as weak AI and strong AI (Searle, 1980), are often employed to distinguish between varying levels of machine intelligence (Russell and Norvig, 2010). While machine learning (ML) is often used interchangeably with AI, they are not identical. Machine learning is a subset of AI and denotes a set of techniques for solving data-related problems without explicit programming (Kühl et al., 2020). In the context of this paper, we define AI as "*the ability of a machine to learn from experience, adjust to new inputs and perform human-like tasks*" (Duan et al., 2019, p. 63). In this paper the term AI encompasses both rule-based and machine learning techniques (Russell and Norvig, 2010).

Also, there are many definitions of human-centric AI (e.g., Wilkens et al., 2021a,b). Our paper contextualizes human-centric AI as AI tools that prioritize and enhance the human experience by making them more intuitive, empathetic, and aligned with human values and needs. Human-centric AI tools understand and respond to human emotions, enabling natural and empathetic interactions, and respect ethical and social considerations in decision-making processes (Del Giudice et al., 2023). One of the challenges in humanizing AI is that there is no universally accepted approach that guides the best practice for design and use of AI tools. The development of human-centric AI should balance human well-being with technical efficiencies (Bingley et al., 2023). We believe that the concept of humanizing AI should be approached from multiple interconnected perspectives to bridge the existing gaps between humans and machines, which is currently lacking in the field (e.g., Han et al., 2021). In a narrow definition, and in the context of this paper, humanizing AI involves the creation and utilization of AI tools that: (i) enhance human potential, build trust, and minimize fear (ii) can interact with humans in a natural, human-like manner, and (iii) can process information during these interactions in a manner similar to human cognitive processes (Fenwick and Molnar, 2022). AI evolves over a path of maturity spanning a continuum of contemporary cognitive architectures to more socio-cognitive and cross-domain architectures (e.g., Gupta et al., 2023), and in terms of implementation and human-centricity, needs to be interpreted in the context of place and time (Wilkens et al., 2021a). These advancements can help create AI with more general intelligence and support ongoing efforts to bring humans and machines closer together.

1.2 The evolving role of HRM; a historical overview

It is important to review the evolution of HRM to better understand how the functions, practices, and philosophies within the field change with time to align with management practices and technological developments, and to identify effective HRM practices in an ever-evolving business environment. Identifying the evolving role HRM has played in humanizing the workplace is equally important.

In the evolution of HRM, existing literature identifies four different stages: administrative HR, personnel management, strategic HR, and business partner HRM (e.g., Fombrun et al., 1984; Kaufman, 2007; Wright, 2008; Kim et al., 2021). Administrative

HR is the organization's earliest phase of human resource practices. During this stage, which was most relevant in the early to mid-20th century, HRM primarily focused on administrative and transactional tasks related to compliance and managing the workforce, using paper-based tools, such as manning tables (Mahoney and Deckop, 1986; Hendrickson, 2003). Administrative HR's focus on humanizing the workplace was mainly concentrated on industrial psychology practices for identifying and selecting new hires and other human factor related activities (Münsterberg, 1998). Personnel Management, which gained prominence in the mid-20th century, marked a transition toward a more employee-oriented approach. In this stage, the primary focus shifted from administrative tasks to effectively managing the workforce as an asset. In this stage, various technology tools, such as applicant tracking systems and learning management systems gained popularity to support recruitment and training processes, enhancing employee skills in a more systematic and efficient manner (Kaufman, 2007; Kim et al., 2021). The tenets of humanizing the workplace in this era were based on a behavioral model, emphasizing the importance of understanding how environmental, social, and psychological factors motivate employee behavior and thus productivity. This gave rise to HR practices such as training and development, employee compensation, and communication (e.g., Kaufman, 2015; Armstrong and Taylor, 2020). Strategic HRM emerged as a transformative stage in the evolution of HRM practices to deal with external pressures such as globalization and technological developments, particularly from the late 20th century onwards. It signified a fundamental shift in HR's role within organizations, evolving from a primarily administrative and personnel-focused function to a proactive and strategic partner role integral to achieving organizational goals (e.g., Kaufman, 2007; Kim et al., 2021). The term HRM originated in this time to encompass its multi-faceted nature. During this era, with the emergence of computers and enterprise resource planning (ERP) systems, human resource information systems (HRIS) were used to store and analyze data to increase workflow efficiencies and make data-driven decisions (Hendrickson, 2003). Humanizing the workplace in the strategic HRM phase focuses mainly on enhancing the employer—employee relationship through improved HRM practices and systems for performance management and career planning leading to higher work satisfaction and productivity (Wright, 2008; Kim et al., 2021). Business partner HRM represents the latest evolution in HR practices. In the business partner HRM era, with the rise of the internet at the turn of the century, there is a heightened focus on digital approaches (e-HRM, online HRM, digital HRM) to make more data-informed decisions and create value for the organization (Wright, 2008; Malik et al., 2020). Seeing employees and talent management as a significant source of competitive advantage, enhancing the human experience at work through technology and people-centric approaches like diversity and inclusion become equally important. In this phase, HRM also recognizes the importance of designing and using technology solutions that align with human values and needs (Malik et al., 2020).

With the advent of AI, firms are assessing how they can implement AI technology to enhance efficiency and productivity (Chui et al., 2023). Humanizing the workplace in the digital HRM phase requires an emphasis on using technology to make the organization more human-centric and enhance human values and potential, which, at times, is contrary to efficiency and productivity goals. The AI-driven phase of business partner HRM is a significant turning point in its evolution. Most organizations are unclear on utilizing AI technologies to achieve their people-management and value enhancement goals, raising concerns about AI ethics, compliance, and culture to create a human-centric workplace (Budhwar et al., 2023).

1.3 The role of HRM in the age of AI

Despite a long history of enhancing physical abilities and basic cognitive skills, technology has never been able to augment human intelligence at the workplace and beyond. This limitation is changing now. For the first time, technology is enabling the enhancement of human intelligence (Abbass, 2019) and this creates new challenges for HRM. Advanced digital technologies (such as AI including cutting edge machine learning techniques) transforming many HRM functions and practices further enhancing HRM across a range of activities and departments to enhance operational performance and value creation (Dwivedi et al., 2021). Despite the range of benefits and opportunities AI presents to organizations, the challenges of effectively integrating AI technology into HRM are complex (Tambe et al., 2019; Palos-Sánchez et al., 2022). Moving forward, it is important to review these challenges in a systematic way to overcome these complexities. We therefore provide a structured framework, grouping HRM practices into three specific bundles: people management, culture, and compliance. People-related functions encompass talent acquisition, development, and management, focusing on the workforce's growth and well-being. Compliance-related functions revolve around adhering to legal and ethical standards, ensuring organizations operate within regulatory boundaries, and maintaining fairness and equity. Culture-related functions concentrate on shaping organizational culture, fostering collaboration, and promoting values and behaviors that align with the firm's mission. By categorizing HRM practices into these three groups, we align with the primary domains where HRM professionals exert their influence (e.g., O'Donovan, 2019; Johnson et al., 2022; Ammirato et al., 2023; Prikshat et al., 2023). This categorization provides a comprehensive view of HRM's role in addressing diverse organizational needs, from nurturing human capital to upholding ethics, meeting regulations, and nurturing a cohesive workplace culture. It also emphasizes that HRM is not solely about administration; it is a strategic business partner that influences people, culture, and compliance to drive the organization's success (Sakka et al., 2022). Furthermore, our recommended framework highlights the need for a multi-disciplinary approach to HRM that considers the technical, ethical, and human elements within each category. In the next section, we explore how HRM can play a pivotal role in bridging the gap between humans and machines in the workplace.

2 How HRM can bring humans and machines closer together in the workplace

The adoption of AI within the field of HRM depends on various technological, business, and human factors. Market demands also impact the decision to use AI within HRM design (e.g., Dwivedi et al., 2021; Nguyen et al., 2022). These factors have varying degrees of development, which can propel or constrain AI implementation within the field of HRM. Moreover, the digitization of HRM (including access to quality and unbiased data) also needs to be carefully managed to mitigate risks and ensure alignment with other business functions (e.g., Malik et al., 2022). It is, therefore, important to review AI design and implementation from a trajectory perspective.

In terms of humanizing AI in the workplace, the function of HRM plays a pivotal and varying role in the process of making AI technical solutions in the workplace more human-centric. The aim is to bring humans and machines closer together. Not taking a human-centric approach to AI usage within HRM not only prevents digital transformation efforts and more data-driven decision-making but also jeopardizes more sustainable human resource management in the digital age (e.g., Budhwar et al., 2022) and further advancement toward safe artificial general intelligence (e.g., Everitt, 2019). Recruitment bias, fear of job loss (Frick et al., 2021; Jöhnk et al., 2021; Uren and Edwards, 2023), ineffective human-machine integration (Arslan et al., 2022), human trust in machines (Gillespie et al., 2021), and concerns of privacy (Bodie, 2022) are some of the most common challenges HRM is facing with AI today and will continue to face moving into the future. Addressing the key challenges at each stage of design and implementation not only helps HRM to reposition itself and the value that it helps create for the organization, but also informs AI development and identifies ways to enhance human properties through emerging technologies.

Drawing insights from literature on technology adaptation within HRM (e.g., Kim et al., 2021), and the future outlook of AI technology (Kurzweil, 2005; Abbass, 2019; Silichev et al., 2019; He et al., 2021), the following subsections discuss three phases of AI usage in the workplace: (1) technocratic, (2) integrated, and (3) fully embedded, specifically for people management, culture, and compliance, the challenges faced at each stage in terms of humanizing AI, and which opportunities HRM can capitalize on (Figure 1). The technocratic phase represents an initial stage of AI-HRM integration, where AI is primarily used to automate and enhance specific HRM functions and practices. It is characterized by the application of AI in tasks such as HR planning, recruitment, training, and performance management. The integrated phase represents a more advanced stage where AI and humans work more closely together. It involves integrating AI into daily functions, personalizing employee experiences, and emphasizing collaboration between humans and machines. The fully-embedded phase reflects a more mature and evolved stage of AI adoption, where HRM focuses on managing the interaction between humans and AI in a way that enhances the overall human experience and seeks to create a workplace that reflects the broader societal goal of leveraging technology for the betterment of individuals and communities. These three phases, from technocratic to fully-embedded, are derived based on the evolution of AI technology adoption within the field of HRM. The first two phases are based on recent empirical literature on AI in HRM (e.g., Arslan et al., 2022; Bansal et al., 2023; Bujold et al., 2023). The last phase is our conceptual view, and it represents a logical progression of how AI is integrated into HRM practices and aligns with broader developments in technology adoption and societal goals (e.g., He et al., 2021).

2.1 AI-HRM human-centric orientation: technocratic phase

Modern technologies, such as AI, machine learning, and AR/VR, play an increasingly vital role within the field of HRM supporting and shaping various people management functions and practices (e.g., Bersin and Chamorro-Premuzic, 2019; Malik et al., 2022). Currently, AI-based applications support HRM professionals with HR planning (e.g., Karatop et al., 2015), selection and recruitment (e.g., Torres and Mejia, 2017; Van Esch et al., 2019), training and development (e.g., Sitzmann and Weinhardt, 2019), performance management (e.g., Bankar and Shukla, 2023), influence employee attitudes such as engagement and work satisfaction (e.g., Castellacci and Viñas-Bardolet, 2019), and support employee retention (e.g., Chowdhury et al., 2023b). AI currently supports and provides HRM functions with various benefits ranging from automating mundane tasks and reducing HR-related costs to debiasing hiring processes and leveraging people analytics to make data-driven decisions (e.g., Henkel et al., 2020).

2.1.1 Challenges

Despite obvious efficiency gains AI brings to organizations, human resource departments are facing new pressures associated with balancing these efficiencies and harmonizing human workforces. AI remains a significant source of concern for employees in many organizations (Palos-Sánchez et al., 2022). Recruitment bias, fear of job loss (Frick et al., 2021; Jöhnk et al., 2021; Uren and Edwards, 2023), ineffective human-machine integration (Arslan et al., 2022), human trust in machines (e.g., Gillespie et al., 2021; Budhwar et al., 2022), managers incomplete understanding of AI systems and their impact on employee outcomes (e.g., Castellacci and Viñas-Bardolet, 2019), existing AI regulatory frameworks too broad to address nuances of AI usage within the context of employment (Chowdhury et al., 2023a), data privacy (Bodie, 2022), and lack of human consideration in AI decision-making (Mazarakis et al., 2023) are some of the most common challenges HRM is navigating with AI.

2.1.2 Opportunities for HRM

The critical role HRM can play in making AI-usage more human-centric is first by providing training and development opportunities to decision-makers in the organization on how AI works and how to use AI in a way that benefits employee and

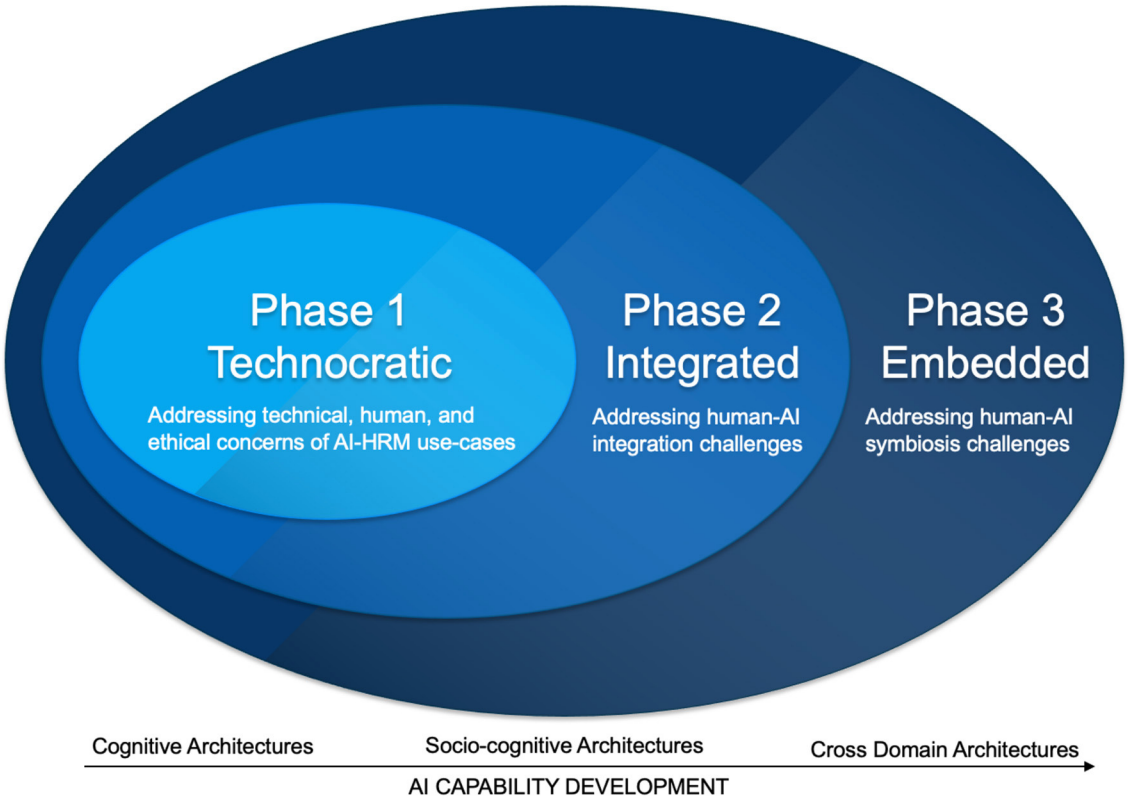

FIGURE 1
The role of HRM in humanizing AI in the workplace.

organizational outcomes (e.g., Arslan et al., 2022; Malik et al., 2022). Second, to address the issues of trust in AI, HRM professionals can play a more active role in addressing concerns about job transformation (Huang et al., 2019), professional identity (e.g., Mirbabaie et al., 2022), AI training (Chowdhury et al., 2023a), and have employees be part of the AI implementation decision-making (e.g., Bankins, 2021; Bankins et al., 2022). Alleviating fears and concerns of employees is critical for AI implementation to succeed in the workplace and to identify more effective ways to implement AI in later stages (e.g., Park et al., 2021). Each of these concerns also affects organizational culture. As more and more machines enter the workforce, replacing human beings, questions are emerging on the changing cultural dynamics within firms (Frangos, 2022; RoŽman et al., 2022; Chowdhury et al., 2023a). In the technocratic stage of AI-HRM implementation and usage it is important to develop and nurture an organizational culture of innovation (Fountaine et al., 2019; Pumplun et al., 2019; Ransbotham et al., 2021), collaboration (Fountaine et al., 2019), and effective change management (Pumplun et al., 2019). From a compliance perspective, firms must start with developing their AI policy to comply with the current high-level guidelines of human-centric AI regulations (e.g., de Laat, 2021). AI policies serve as a critical foundation to support AI implementation and usage within the organization, maintain ethical standards, and develop trust with internal and external stakeholders (Sjödin et al., 2021). Finally, HRM can also work as an interface between developers and employees to help address the lack of human consideration when AI makes critical decisions about hiring, firing, and reward allocation (e.g., Malik et al., 2023).

2.2 AI-HRM human-centric orientation: human-AI integration phase

Human-AI integration can happen to varying degrees. To date, most human-AI integration focuses on the co-existence of humans together with AI, where humans and AI perform as separate entities. Recent AI developments focus more on human-AI integration, where humans and machines make decisions together (e.g., Einola and Khoreva, 2023). This is often referred to as human-in-the-loop (HITL) (e.g., Monarch and Munro, 2021). In phase two, HRM practices focus on bringing humans and machines closer together by integrating AI more into daily functions of employees (e.g., Rydén and El Sawy, 2022), personalizing employee experiences and learning journeys (e.g., Bulut and Özlem, 2023), and identifying and leveraging human-AI interaction mechanisms in the workplace (e.g., Budhwar et al., 2022; Herrmann and Pfeiffer, 2023). When we look at empirical survey data, high AI performer firms, defined as "*organizations that attribute at least 20 percent of their EBIT to AI adoption*" (Chui et al., 2023, p. 8), already distinguish themselves by integrating AI deeply into their operations, leveraging it not just for cost reduction but to enhance HRM functions and organization design. This comprehensive use

of AI in enhancing organizational design and creating new value propositions sets high AI performer firms apart, demonstrating a more integrated and strategic application of AI within their organizations (Chui et al., 2023). As human and machine systems and processes become more integrated in phase two, organizational culture management will evolve as well. Leadership style shifts are most likely to occur as a result of changing employee dynamics influenced by AI implementation (Peifer et al., 2022). In phase two, firms move beyond high-level regulations to anticipate and implement more prescriptive guidelines and controls. This phase will be characterized by meeting not only current regulations but preparing for future regulations designed to address AI's unique challenges (e.g., Hadfield and Clark, 2023). Compliance also plays a stronger role in responsible human-computer interaction (HCI) design and human-computer responsibilities and liabilities (e.g., Rakova et al., 2021).

2.2.1 Challenges

Human-AI integration phase faces unique challenges. Some of the challenges HRM will face in the integration phase are role and job design challenges (e.g., Sampson, 2021), HCI design challenges (e.g., Arslan et al., 2022), human and AI cross-functional team issues (e.g., Klien et al., 2004; Arslan et al., 2022), responsible design (e.g., Bankins, 2021), ethical concerns in terms of decision-making (e.g., Flathmann et al., 2021), cultural differences (Herrmann and Pfeiffer, 2023), and appropriate oversight and governance (e.g., Wu et al., 2020). The main challenges HRM faces in phase two are centered around employee up-skilling and re-skilling, AI solution design and integration challenges, and delineation of responsibility between humans and machines.

2.2.2 Opportunities for HRM

To help address these issues, HRM professionals first can focus training efforts on augmenting existing skills using AI tools and applications so that employees feel more comfortable working with AI technology and making decisions together (e.g., Arslan et al., 2022). Second, HRM continues to work with AI application developers to make sure integrated AI usage is user-friendly, intuitive, explainable, and responsible. Third, study the human-AI interactive mechanisms that amplify human skills and develop guidelines for human-AI collaboration and integration (e.g., Budhwar et al., 2022; Berretta et al., 2023; Hu and Wu, 2023). These efforts to take a human-centric approach to learning and development can motivate employees to learn how to work with new technologies and be more willing to transform with the organization (e.g., Beichter and Kaiser, 2023). Integrated AI tools can also augment human capabilities through a HITL approach in which humans participate in the algorithmic decision-making process, improving the explainability of decision outcomes and human acceptance of algorithm-based decisions (Mosqueira-Rey et al., 2023). As technology advances and moves more into socio-cognitive architecture models, more advanced HITL setups will emerge (e.g., Gupta et al., 2023; Mosqueira-Rey et al., 2023). Finally, anticipating ongoing changes to regulation, including but not limited to anticipated compliance verification requirements, organizations at this stage stay committed to building continuous learning and adaptation mechanisms to minimize liabilities and unethical AI usage in the workplace (e.g., Kulkarni et al., 2021; Wiehler, 2022; Grabowicz et al., 2023; Hu and Wu, 2023).

2.3 AI-HRM human-centric orientation: fully-embedded AI phase

The advancement of new AI architectures (moving more toward cross-domain intelligence) and human-computer interaction, together with operationalizing human-AI collaboration in the workplace, starts a new phase in the AI-driven organization. In the fully-embedded phase, AI is more intelligent and less artificial, becoming an imperative within organizations for creating and capturing value. Once the AI-driven organization is fully operational and traditional HRM functions and practices are automated, the role of HRM focuses less on integration and emphasizes more on employee experience and organizational effectiveness, ensuring that they are in line with human-centric principles and ethical standards (e.g., Seidl, 2022). In the fully AI-embedded phase, the functions and processes of HRM are very different than in previous stages. The function of HRM becomes more strategic and human-centered and will focus more on managing organizational and algorithmic behavior to help the organization meet rapidly changing needs (e.g., Langer and König, 2023; Rodgers et al., 2023). The role of HRM includes the management of human resources and technology together due to its increased symbiotic relationship. In the fully-embedded AI phase, HRM becomes an even more multi-disciplinary function, working together with behavioral data scientists, psychologists, and technologists (Fenwick and Molnar, 2022), we therefore propose HRM to reposition itself to Human Technology Resource Management (HTRM).

2.3.1 Challenges

Technology and human resources are both equally important, and the challenge for HRM is to build (or, keep building) a symbiotic relationship between humans and machines. Besides the ongoing focus for re-skilling and job design, challenges could be employee resistance to fully automated AI-HRM (e.g., Brock and von Wangenheim, 2019; Frick et al., 2021), bias and fairness checks (e.g., Zhuo et al., 2023), maintaining human-centricity and purpose-driven approaches (e.g., Cappelli and Rogovsky, 2023), and complex human issues and well-being, such as digital divide and mental health issues (e.g., Khogali and Mekid, 2023). Most of the challenges in phase three center around human well-being, performance optimization, exception handling, and ethics. Increased automation is known to lead to more stress and anxiety in the workplace amongst other psycho-social risks (e.g., Cefaliello, 2021). As AI-powered tools and processes become more "intelligent," human employees can fear AI and harbor job insecurities and unfair treatment.

2.3.2 Opportunities for HRM

HRM could address these issues from a human-centric approach by ensuring humans are put at the center of AI-HRM development (e.g., Mazarakis et al., 2023). Looking ahead to industry 5.0 (e.g., Coelho et al., 2023), there is a greater focus on the human aspect within organizations aiming to find more sustainable and resilient ways to bring humans and machines together thus rethinking how value is created in today's world (e.g., Del Giudice et al., 2023; Pizoń and Gola, 2023). In phase three, new perspectives of human-AI integration at work are extending to neural integration, where AI tools are embedded into humans (e.g., mind-controlled machines, neurolinks, intelligent prostheses) to enhance human capabilities or human cells are used in bioengineering for the development of organoid intelligence (e.g., Morales Pantoja et al., 2023). With the emergence of advanced integrated human-AI tools and interfaces, we predict that HRM will continue to focus on developing strict adherence to ethical rules (e.g., Pflanzer et al., 2023). The HRM community will also influence regulators to enforce more human-centric policies. Emphasizing the importance of culture in mitigating employee resistance remains a pressing concern for HRM in the future (Ransbotham et al., 2021), as is addressing issues concerning centralized power with the AI-embedded organization (e.g., Einola and Khoreva, 2023). This approach not only fosters ethical AI but also distinguishes organizations as stewards of technology that enhances, rather than diminishes, the human experience.

3 Conclusion

In the age of AI, the role of HRM professionals in organizations continues to evolve. AI technologies are increasingly being implemented in organizations to enhance HRM across a range of activities and departments to support operational performance and value creation. A growing body of evidence highlights the benefits AI brings to the field of HRM. Despite the growing interest in AI-HRM scholarship, the focus on human-AI interaction at work and AI-based technologies for HRM is limited and fragmented. Moreover, the lack of human considerations in HRM tech design and deployment can hamper AI digital transformation efforts and jeopardize more sustainable human resource practices in the digital age and even advancements toward safe artificial general intelligence. To provide a structured framework for reviewing these challenges, and based on existing literature (e.g., Ammirato et al., 2023; Prikshat et al., 2023), we grouped HRM practices into three specific bundles: people management, culture, and compliance. By categorizing HRM functions into these three groups, we align with the primary domains where HRM support is most needed in the age of AI integration in the workplace.

Our paper underscores the dynamic evolution of HRM in the era of AI, emphasizing its central role in orchestrating the integrated and symbiotic relationship between humans and machines within organizations. The lack of understanding in implementing AI in a human-centric way highlights the need for a practical approach that goes beyond merely humanizing AI. HRM plays a pivotal role in this area seeing its human-centric focus in the value creation process of organizations and its strategic position within management practice to enhance organizational effectiveness. We propose adopting a multi-disciplinary, human-centric, and integrated approach that can address the current concerns and fears surrounding AI development and deployment in the workplace. AI evolves over a path of maturity spanning a continuum of contemporary cognitive architectures to more socio-cognitive and cross-domain architectures (e.g., Gupta et al., 2023), and in terms of implementation and human-centricity, needs to be interpreted in the context of place and time (Wilkens et al., 2021a). This paper, therefore, categorizes the AI-HRM journey into technocratic, human-AI integration, and fully-embedded AI phases, each presenting unique challenges and opportunities. The benefit of this approach is that it allows organizations to evaluate at which stage of AI implementation and usage they find themselves and the critical role HRM can play in advancing digital transformation efforts and human-AI integration. In our paper, we also anticipate the emergence of advanced human-AI integration paradigms, such as neural integration, emphasizing HRM's role in ensuring ethical, responsible, and fair practices. By looking at the issue from culture, compliance, and people management, our framework not only paves a roadmap toward human-centric AI, but also distinguishes organizations as stewards of technology that enhances, rather than diminishes, the human experience and potential. The paper serves as a forward-looking guide for HRM practitioners, policymakers, and researchers seeking to navigate the transformative landscape of AI in HRM while upholding ethical principles and fostering a future where AI and humans symbiotically co-exist in the workplace.

Author contributions

AF: Writing—original draft, Writing—review & editing. GM: Writing—original draft, Writing—review & editing. PF: Writing—original draft, Writing—review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Abbass, H. A. (2019). Social integration of artificial intelligence: functions, automation allocation logic and human-autonomy trust. *Cognit. Comput.* 11, 159–171. doi: 10.1007/s12559-018-9619-0

Afiouni, R. (2019). *Organizational Learning in the Rise of Machine Learning (2019)*. ICIS 2019 Proceedings, Munich. Available online at: https://aisel.aisnet.org/icis2019/business_models/business_models/2 (accessed November 1, 2023).

Agrawal, A., Gans, J., and Goldfarb, A. (2017). What to expect from artificial intelligence. *MIT Sloan Manag. Rev.* 58, 1. Available online at: https://sloanreview.mit.edu/article/what-to-expect-from-artificial-intelligence/ (accessed November 1, 2023).

Ammirato, S., Felicetti, A. M., Linzalone, R., Corvello, V., and Kumar, S. (2023). Still our most important asset: a systematic review on human resource management in the midst of the fourth industrial revolution. *J. Innov. Knowl.* 8, 100403. doi: 10.1016/j.jik.2023.100403

Armstrong, M., and Taylor, S. (2020). *Armstrong's Handbook of Human Resource Management Practice*. London: Kogan Page Publishers.

Arslan, A., Cooper, C., Khan, Z., Golgeci, I., and Ali, I. (2022). Artificial intelligence and human workers interaction at team level: a conceptual assessment of the challenges and potential HRM strategies. *Int. J. Manpow.* 43, 75–88. doi: 10.1108/IJM-01-2021-0052

Bankar, S., and Shukla, K. (2023). "Performance management and artificial intelligence: a futuristic conceptual framework," in *Contemporary Studies of Risks in Emerging Technology, Part B (Emerald Studies in Finance, Insurance, and Risk Management)*, eds S. Grima, K. Sood, and E. Özen (Leeds: Emerald Publishing Limited), 341–360.

Bankins, S. (2021). The ethical use of artificial intelligence in human resource management: a decision-making framework. *Ethics Inf. Technol.* 23, 841–854. doi: 10.1007/s10676-021-09619-6

Bankins, S., Formosa, P., Griep, Y., and Richards, D. (2022). AI decision making with dignity? Contrasting workers' justice perceptions of human and AI decision making in a human resource management context. *Inf. Syst. Front.* 24, 857–875. doi: 10.1007/s10796-021-10223-8

Bansal, A., Panchal, T., Jabeen, F., Mangla, S. K., and Singh, G. (2023). A study of human resource digital transformation (HRDT): a phenomenon of innovation capability led by digital and individual factors. *J. Bus. Res.* 157, 113611. doi: 10.1016/j.jbusres.2022.113611

Beer, M., Boselie, P., and Brewster, C. (2015). Back to the future: implications for the field of HRM of the multistakeholder perspective proposed 30 years ago. *Hum. Resour. Manage.* 54, 3. 427–438. doi: 10.1002/hrm.21726

Beichter, T., and Kaiser, M. (2023). "The future of upskilling: human- and technology-centered," in *Proceedings of the 14th International Multi-Conference on Complexity, Informatics and Cybernetics: IMCIC 2023*, eds N. Callaos, S. Hashimoto, N. Lace, B. Sánchez, and M. Savoie (Orlando, FL: International Institute of Informatics and Cybernetics), 190–193. doi: 10.54808/IMCIC2023.01.190

Berretta, S., Tausch, A., Ontrup, G., Gilles, B., Peifer, C., Kluge, A., et al. (2023). Defining human-AI teaming the human-centered way: a scoping review and network analysis. *Front. Artif. Intell.* 6, 1250725. doi: 10.3389/frai.2023.1250725

Bersin, J., and Chamorro-Premuzic, T. (2019). New ways to gauge talent and potential. *MIT Sloan Manag. Rev.* 60, 1. Available online at: https://sloanreview.mit.edu/article/new-ways-to-gauge-talent-and-potential/ (accessed October 28, 2023).

Bingley, W. J., Curtis, C., Lockey, S., Bialkowski, A., Gillespie, N., Haslam, S. A., et al. (2023). Where is the human in human-centered AI? Insights from developer priorities and user experiences. *Comput. Hum. Behav.* 141, 107617. doi: 10.1016/j.chb.2022.107617

Bodie, M. T. (2022). The law of employee data: privacy, property, governance. *Indiana Law J.* 97, 707–753. Available online at: https://www.repository.law.indiana.edu/ilj/vol97/iss2/7 (accessed October 28, 2023).

Boselie, P., Van Harten, J., and Veld, M. (2021). A human resource management review on public management and public administration research: stop right there… before we go any further…. *Public Manag. Rev.* 23, 483–500. doi: 10.1080/14719037.2019.1695880

Bresciani, S., Ciampi, F., Meli, F., and Ferraris, A. (2021). Using big data for co-innovation processes: Mapping the field of data-driven innovation, proposing theoretical developments and providing a research agenda. *Int. J. Inf. Manage.* 60, 102347. doi: 10.1016/j.ijinfomgt.2021.102347

Brock, J. K. U., and von Wangenheim, F. (2019). Demystifying AI: what digital transformation leaders can teach you about realistic artificial intelligence. *Calif. Manage. Rev.* 61, 4. 110-134. doi: 10.1177/1536504219865226

Budhwar, P., Chowdhury, S., Wood, G., Aguinis, H., Bamber, G. J., Beltran, J. R., et al. (2023). Human resource management in the age of generative artificial intelligence: perspectives and research directions on ChatGPT. *Hum. Resour. Manag. J.* 33, 606–659. doi: 10.1111/1748-8583.12524

Budhwar, P., Malik, A., De Silva, M. T., and Thevisuthan, P. (2022). Artificial intelligence–challenges and opportunities for international HRM: a review and research agenda. *Int. J. Hum. Resour. Manag.* 33, 1065–1097. doi: 10.1080/09585192.2022.2035161

Bujold, A., Roberge-Maltais, I., Parent-Rocheleau, X., Boasen, J., Sénécal, S., Léger, P. M., et al. (2023). Responsible artificial intelligence in human resources management: a review of the empirical literature. *AI Ethics* 1–16. doi: 10.1007/s43681-023-00325-1

Bulut, A., and Özlem, D. (2023). The effect of industry 4.0 and artificial intelligence on human resource management. *Int. J. East. Anatol. Sci. Eng. Des.* 5, 2. 143–166. doi: 10.47898/ijeased.1306881

Cappelli, P., and Rogovsky, N. G. (2023). *Artificial Intelligence in Human Resource Management: A Challenge for the Human-centred Agenda? (No. 95)*. ILO Working Paper. (Geneva: ILO).

Castellacci, F., and Viñas-Bardolet, C. (2019). Internet use and job satisfaction. *Comput. Hum. Behav.* 90, 141–152. doi: 10.1016/j.chb.2018.09.001

Cefaliello, A. (2021). *Psychosocial Risks in Europe: National Examples as Inspiration for a Future Directive*. ETUI Research Paper-Policy Briefs. (Brussels: ETUI).

Chowdhury, S., Dey, P., Joel-Edgar, S., Bhattacharya, S., Rodriguez-Espindola, O., Abadie, A., et al. (2023a). Unlocking the value of artificial intelligence in human resource management through AI capability framework. *Hum. Resour. Manag. Rev.* 33, 100899. doi: 10.1016/j.hrmr.2022.100899

Chowdhury, S., Joel-Edgar, S., Dey, P. K., Bhattacharya, S., and Kharlamov, A. (2023b). Embedding transparency in artificial intelligence machine learning models: managerial implications on predicting and explaining employee turnover. *Int. J. Hum. Resour. Manage.* 34, 2732-2764. doi: 10.1080/09585192.2022.2066981

Chui, M., Yee, L., Hall, B., and Singla, A. (2023). *The State of AI in 2023: Generative AI's Breakout Year*. Brisbane: McKinsey Global Publishing.

Coelho, P., Bessa, C., Landeck, J., and Silva, C. (2023). Industry 5.0: the arising of a concept. *Procedia Comput. Sci.* 217, 1137–1144. doi: 10.1016/j.procs.2022.12.312

de Laat, P. B. (2021). Companies committed to responsible AI: from principles towards implementation and regulation? *Philos. Technol.* 34, 1135–1193. doi: 10.1007/s13347-021-00474-3

De Stefano, V., and Wouters, M. (2022). *AI and Digital Tools in Workplace Management and Evaluation: An Assessment of the EU's Legal Framework*. Brussels: European Parliamentary Research Service.

Del Giudice, M., Scuotto, V., Orlando, B., and Mustilli, M. (2023). Toward the human–centered approach. A revised model of individual acceptance of AI. *Hum. Resour. Manag. Rev.* 33, 100856. doi: 10.1016/j.hrmr.2021.100856

DiClaudio, M. (2019). People analytics and the rise of HR: how data, analytics and emerging technology can transform human resources (HR) into a profit center. *Strategic HR Rev.* 18, 42–46. doi: 10.1108/SHR-11-2018-0096

Duan, Y., Edwards, J. S., and Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of Big Data – evolution, challenges, and research agenda. *Int. J. Inf. Manage.* 48, 63–71. doi: 10.1016/j.ijinfomgt.2019.01.021

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., et al. (2021). Artificial Intelligence (AI): multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice, and policy. *Int. J. Inf. Manage.* 57, 101994. doi: 10.1016/j.ijinfomgt.2019.08.002

Einola, K., and Khoreva, V. (2023). Best friend or broken tool? Exploring the co-existence of humans and artificial intelligence in the workplace ecosystem. *Hum. Resour. Manage.* 62, 117–135. doi: 10.1002/hrm.22147

Everitt, T. (2019). *Towards Safe Artificial General Intelligence* [Doctoral dissertation]. The Australian National University (Australia). Available online at: https://www.proquest.com/docview/2353149751 (accessed October 28, 2023).

Fenwick, A., and Molnar, G. (2022). The importance of humanizing AI: using a behavioral lens to bridge the gaps between humans and machines. *Discov. Artif. Intell.* 2, 14. doi: 10.1007/s44163-022-00030-8

Flathmann, C., Schelble, B. G., Zhang, R., and McNeese, N. J. (2021). "Modeling and guiding the creation of ethical human-AI teams," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY: ACM), 469–479. doi: 10.1145/3461702.3462573

Fombrun, C. J., Tichy, N. M., and Devanna, M. A. (1984). *Strategic Human Resource Management*. New York, NY: Wiley.

Fountaine, T., McCarthy, B., and Saleh, T. (2019). Building the AI-powered organization. *Harv. Bus. Rev.* 97, 62–73. Available online at: https://hbr.org/2019/07/building-the-ai-powered-organization (accessed November 1, 2023).

Frangos, P. (2022). An integrative literature review on leadership and organizational readiness for AI. *Eur. Conf. Impact Artif. Intell. Robot.* 4, 145–152. Available online at: https://papers.academic-conferences.org/index.php/ecair/article/view/834/914 (accessed November 1, 2023).

Frick, N. R., Mirbabaie, M., Stieglitz, S., and Salomon, J. (2021). Maneuvering through the stormy seas of digital transformation: the impact of empowering leadership on the AI readiness of enterprises. *J. Decis. Syst.* 30, 235–258. doi: 10.1080/12460125.2020.1870065

Gillespie, N., Lockey, S., and Curtis, C. (2021). Trust in Artificial Intelligence*: A Five Country Study*. St Lucia, QLD: The University of Queensland and KPMG Australia. doi: 10.14264/e34bfa3

Grabowicz, P., Perello, N., and Zick, Y. (2023). Towards an AI accountability policy. *arXiv* [preprint]. doi: 10.48550/arXiv.2307.13658

Guenole, N., and Feinzig, S. (2018). *The Business Case for AI in HR. With Insights and Tips on Getting Started*. Armonk, NY: IBM Smarter Workforce Institute; IBM Corporation. Available online at: https://www.ibm.com/downloads/cas/A5YLEPBR (accessed October 26, 2023).

Gupta, P., Nguyen, T. N., Gonzalez, C., and Woolley, A. W. (2023). Fostering collective intelligence in human–AI collaboration: laying the groundwork for COHUMAIN. *Top. Cogn. Sci.*. 1–28. doi: 10.1111/tops.12679

Hadfield, G. K., and Clark, J. (2023). Regulatory markets: the future of AI governance. *arXiv* [preprint]. doi: 10.48550/arXiv.2304.04914

Han, S., Kelly, E., Nikou, S., and Svee, E. O. (2021). Aligning artificial intelligence with human values: reflections from a phenomenological perspective. *AI Soc.* 37, 1383–1395. doi: 10.1007/s00146-021-01247-4

He, H., Gray, J., Cangelosi, A., Meng, Q., McGinnity, T. M., Mehnen, J., et al. (2021). The challenges and opportunities of human-centered AI for trustworthy robots and autonomous systems. *IEEE Trans. Cogn. Dev. Syst.* 14, 1398–1412. doi: 10.1109/TCDS.2021.3132282

Hendrickson, A. R. (2003). Human resource information systems: backbone technology of contemporary human resources. *J. Labor Res.* 24, 381. doi: 10.1007/s12122-003-1002-5

Henkel, A. P., Caić, M., Blaurock, M., and Okan, M. (2020). Robotic transformative service research: deploying social robots for consumer well-being during COVID-19 and beyond. *J. Serv. Manag.* 31, 1131–1148. doi: 10.1108/JOSM-05-2020-0145

Herrmann, T., and Pfeiffer, S. (2023). Keeping the organization in the loop: a socio-technical extension of human-centered artificial intelligence. *AI Soc.* 38, 1523–1542. doi: 10.1007/s00146-022-01391-5

Hu, B., and Wu, Y. (2023). AI-based compliance automation in commercial bank: how the silicon valley bank provided a cautionary tale for future integration. *Int. Res. Econ. Finan.* 7, 13. doi: 10.20849/iref.v7i1.1356

Huang, M. H., Rust, R., and Maksimovic, V. (2019). The feeling economy: managing in the next generation of artificial intelligence (AI). *Calif. Manage. Rev.* 61, 43–65. doi: 10.1177/0008125619863436

Jöhnk, J., Weibert, M., and Wyrtki, K. (2021). Ready or not, AI comes— an interview study of organizational AI readiness factors. *Bus. Inf. Syst. Eng.* 63, 5–20. doi: 10.1007/s12599-020-00676-7

Johnson, B. A. M., Coggburn, J. D., and Llorens, J. J. (2022). Artificial intelligence and public human resource management: questions for research and practice. *Public Pers. Manage.* 51, 538–562. doi: 10.1177/00910260221126498

Karatop, B., Kubat, C., and Uygun, Ö. (2015). Talent management in manufacturing system using fuzzy logic approach. *Comput. Ind. Eng.* 86, 127–136. doi: 10.1016/j.cie.2014.09.015

Kaufman, B. E. (2007). "The development of HRM in historical and international perspective," in *The Oxford Handbook of Human Resource Management*, eds P. Boxall, J. Purcell, and P. M. Wright (Oxford: Oxford University Press), 19–47. doi: 10.1093/oxfordhb/9780199547029.003.0002

Kaufman, B. E. (2015). Evolution of strategic HRM as seen through two founding books: a 30th anniversary perspective on development of the field. *Hum. Resour. Manage.* 54, 389–407. doi: 10.1002/hrm.21720

Khogali, H. O., and Mekid, S. (2023). The blended future of automation and AI: examining some long-term societal and ethical impact features. *Technol. Soc.* 73, 102232. doi: 10.1016/j.techsoc.2023.102232

Kim, S., Wang, Y., and Boon, C. (2021). Sixty years of research on technology and human resource management: looking back and looking forward. *Hum. Resour. Manage.* 60, 229–247. doi: 10.1002/hrm.22049

Klien, G., Woods, D. D., Bradshaw, J. M., Hoffman, R. R., and Feltovich, P. J. (2004). Ten challenges for making automation a "team player" in joint human-agent activity. *IEEE Intell. Syst.* 19, 91–95. doi: 10.1109/MIS.2004.74

Kühl, N., Goutier, M., Hirt, R., and Satzger, G. (2020). Machine learning in artificial intelligence: towards a common understanding. *arXiv* [preprint]. doi: 10.48550/arXiv.2004.04686

Kulkarni, V., Sunkle, S., Kholkar, D., Roychoudhury, S., Kumar, R., Raghunandan, M., et al. (2021). Toward automated regulatory compliance. *CSI Trans. ICT* 9, 95–104. doi: 10.1007/s40012-021-00329-4

Kurzweil, R. (2005). *The Singularity Is Near: When Humans Transcend Biology*. New York, NY: Penguin Books.

Langer, M., and König, C. J. (2023). Introducing a multi-stakeholder perspective on opacity, transparency and strategies to reduce opacity in algorithm-based human resource management. *Hum. Resour. Manag. Rev.* 33, 100881. doi: 10.1016/j.hrmr.2021.100881

Lee, J., Suh, T., Roy, D., and Baucus, M. (2019). Emerging technology and business model innovation: the case of artificial intelligence. *J. Open Innov.: Technol. Mark. Complex* 5, 44. doi: 10.3390/joitmc5030044

Mahoney, T. A., and Deckop, J. R. (1986). Evolution of concept and practice in personnel administration/human resource management (PA/HRM). *J. Manage.* 12, 223–241. doi: 10.1177/014920638601200206

Malik, A., Budhwar, P., and Kazmi, B. A. (2022). Artificial intelligence (AI)-assisted HRM: towards an extended strategic framework. *Hum. Resour. Manag. Rev.* 33, 100940. doi: 10.1016/j.hrmr.2022.100940

Malik, A., Budhwar, P., and Mohan, H. NR, S. (2023). Employee experience–the missing link for engaging employees: insights from an MNE's AI-based HR ecosystem. *Hum. Resour. Manage.* 62, 97–115. doi: 10.1002/hrm.22133

Malik, A., Srikanth, N. R., and Budhwar, P. (2020). "Digitisation, artificial intelligence (AI) and HRM," in *Human Resource Management: Strategic and International Perspectives*, eds J. Crawshaw, P. Budhwar, and A. Davis (London: SAGE Publications), 88–111.

Mazarakis, A., Bernhard-Skala, C., Braun, M., and Peters, I. (2023). What is critical for human-centered AI at work? -towards an interdisciplinary theory. *Front. Artif. Intell.* 6, 1257057. doi: 10.3389/frai.2023.1257057

Mikalef, P., and Gupta, M. (2021). Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance. *Inf. Manag.* 58, 103434. doi: 10.1016/j.im.2021.103434

Mirbabaie, M., Brünker, F., Möllmann, N. R., and Stieglitz, S. (2022). The rise of artificial intelligence–understanding the AI identity threat at the workplace. *Electron. Mark.* 32, 73–99. doi: 10.1007/s12525-021-00496-x

Monarch, R., and Munro, R. (2021). *Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-centered AI*. New York, NY: Simon and Schuster.

Morales Pantoja, I. E., Smirnova, L., Muotri, A. R., Wahlin, K. J., Kahn, J., Boyd, J. L., et al. (2023). First organoid intelligence (OI) workshop to form an OI community. *Front. Artif. Intell.* 6, 1116870. doi: 10.3389/frai.2023.1116870

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., and Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: a state of the art. *Artif. Intell. Rev.* 56, 3005–3054. doi: 10.1007/s10462-022-10246-w

Münsterberg, H. (1998). *Psychology and Industrial Efficiency* (R. H. Wozniak, ed.). Bristol: Thoemmes Press.

Nguyen, T. M., Malik, A., and Budhwar, P. (2022). Knowledge hiding in organizational crisis: the moderating role of leadership. *J. Bus. Res.* 139, 161–172. doi: 10.1016/j.jbusres.2021.09.026

O'Donovan, D. (2019). "HRM in the organization: an overview," in *Management Science: Foundations and Innovations*, eds C. Machado, and J. Davim (Cham: Springer), 75–110. doi: 10.1007/978-3-030-13229-3_4

Palos-Sánchez, P. R., Baena-Luna, P., Badicu, A., and Infante-Moro, J. C. (2022). Artificial intelligence and human resources management: a bibliometric analysis. *Appl. Artif. Intell.* 36, 2145631. doi: 10.1080/08839514.2022.2145631

Park, H., Ahn, D., Hosanagar, K., and Lee, J. (2021). "Human-AI interaction in human resource management: understanding why employees resist algorithmic evaluation at workplaces and how to mitigate burdens," in Proceedings of the *2021 CHI Conference on Human Factors in Computing Systems* (New York, NY: ACM), 1–15. doi: 10.1145/3411764.3445304

Peifer, Y., Jeske, T., and Hille, S. (2022). Artificial intelligence and its impact on leaders and leadership. *Procedia Comput. Sci.* 200, 1024–1030. doi: 10.1016/j.procs.2022.01.301

Pflanzer, M., Traylor, Z., Lyons, J. B., Dubljević, V., and Nam, C. S. (2023). Ethics in human–AI teaming: principles and perspectives. *AI Ethics* 3, 917–935. doi: 10.1007/s43681-022-00214-z

Pizoń, J., and Gola, A. (2023). Human–machine relationship—perspective and future roadmap for Industry 5.0 solutions. *Machines* 11, 203. doi: 10.3390/machines11020203

Prikshat, V., Malik, A., and Budhwar, P. (2023). AI-augmented HRM: Antecedents, assimilation and multilevel consequences. *Hum. Resour. Manag. Rev.* 33, 100860. doi: 10.1016/j.hrmr.2021.100860

Pumplun, L., Tauchert, C., and Heidt, M. (2019). "A new organizational chassis for artificial intelligence-exploring organizational readiness factors," in *Volume 27, European Conference on Information Systems (ECIS'2019)*. Stockholm.

Rakova, B., Yang, J., Cramer, H., and Chowdhury, R. (2021). Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proc. ACM Hum.-Comput. Interact.* 5(CSCW1), 1–23. doi: 10.1145/3449081

Ransbotham, S., Candelon, F., and Kiron, D. LaFountain, B., Khodabandeh, S. (2021). The Cultural Benefits of Artificial Intelligence in the Enterprise. *Group*. Available online at: https://web-assets.bcg.com/85/90/95939185404cbd901aba0d54f1d7/the-cultural-benefits-of-artificial-intelligence-in-the-enterprise-r.pdf (accessed November 1, 2023).

Rathi, R. A. (2018). Artificial intelligence and the future of hr practices. *Int. J. Appl. Res.* 4, 113–116. Available online at: https://www.allresearchjournal.com/archives/2018/vol4issue6/PartB/4-6-18-226.pdf (accessed November 1, 2023).

Rodgers, W., Murray, J. M., Stefanidis, A., Degbey, W. Y., and Tarba, S. Y. (2023). An artificial intelligence algorithmic approach to ethical decision-making in human resource management processes. *Hum. Resour. Manag. Rev.* 33, 100925. doi: 10.1016/j.hrmr.2022.100925

RoŽman, M., Oreški, D., and Tominc, P. (2022). Integrating artificial intelligence into a talent management model to increase the work engagement and performance of enterprises. *Front. Psychol.* 13, 1014434. doi: 10.3389/fpsyg.2022.1014434

Russell, S. J., and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. London: Pearson.

Rydén, P., and El Sawy, O. (2022). "Real-time management: when AI goes fast and flow," in *Platforms and Artificial Intelligence: The Next Generation of Competences*, ed A. Bounfour (Cham: Springer International Publishing), 225–243.

Sakka, F., El Maknouzi, M. E. H., and Sadok, H. (2022). Human resource management in the era of artificial intelligence: future HR work practices, anticipated skill set, financial and legal implications. *Acad. Strateg. Manag. J.* 21, 1–14. Available online at: https://www.abacademies.org/articles/human-resource-management-in-the-era-of-artificial-intelligence-future-hr-work-practices-anticipated-skill-set-financial-and-legal-13536.html (accessed November 1, 2023).

Sampson, S. E. (2021). A strategic framework for task automation in professional services. *J. Serv. Res.* 24, 122–140. doi: 10.1177/1094670520940407

Schmidt, R., Zimmermann, A., Möhring, M., and Keller, B. (2020). "Value creation in connectionist artificial intelligence-a research agenda," in *AMCIS 2020 Proceedings-Advancings in Information Systems Research: August 10–14, 2020* (Atlanta, GA: AMCIS), 1–10.

Searle, J. R. (1980). Minds, brains and programs. *Behav. Brain Sci.* 3, 417–457. doi: 10.1017/S0140525X00005756

Seidl, M. (2022). "Corporate digital responsibility: stimulating human-centric innovation and building trust in the digital world," in *Liquid Legal - Humanization and the Law*, eds K. Jacob, D. Schindler, R. Strathausen, and B. Waltl (Cham: Springer International Publishing), 55–81. doi: 10.1007/978-3-031-14240-6_4

Silichev, D., Volobuev, A., and Kuzina, E. (2019). "Artificial intelligence and the future of the mankind," in *Ubiquitous Computing and the Internet of Things: Prerequisites for the Development of ICT. Studies in Computational Intelligence*, ed. E. Popkova (Cham: Springer International Publishing), 826. doi: 10.1007/978-3-030-13397-9_74

Simon, H. A. (1996). *The Sciences of the Artificial*. Cambridge: MIT press.

Sitzmann, T., and Weinhardt, J. M. (2019). Approaching evaluation from a multilevel perspective: a comprehensive analysis of the indicators of training effectiveness. *Hum. Resour. Manag. Rev.* 29, 253–269. doi: 10.1016/j.hrmr.2017.04.001

Sjödin, D., Parida, V., Palmié, M., and Wincent, J. (2021). How AI capabilities enable business model innovation: scaling AI through co-evolutionary processes and feedback loops. *J. Bus. Res.* 134, 574–587. doi: 10.1016/j.jbusres.2021.05.009

Tambe, P., Cappelli, P., and Yakubovich, V. (2019). Artificial intelligence in human resources management: challenges and a path forward. *Calif. Manage. Rev.* 61, 15–42. doi: 10.1177/0008125619867910

Torres, E. N., and Mejia, C. (2017). Asynchronous video interviews in the hospitality industry: considerations for virtual employee selection. *Int. J. Hosp. Manag.* 61, 4–13. doi: 10.1016/j.ijhm.2016.10.012

Uren, V., and Edwards, J. S. (2023). Technology readiness and the organizational journey towards AI adoption: an empirical study. *Int. J. Inf. Manage.* 68, 102588. doi: 10.1016/j.ijinfomgt.2022.102588

Van Esch, P., Black, J. S., and Ferolie, J. (2019). Marketing AI recruitment: the next phase in job application and selection. *Comput. Human Behav.* 90, 215–222. doi: 10.1016/j.chb.2018.09.009

Wang, P. (2019). On defining artificial intelligence. *J. Artif. Gen. Intell.* 10, 1–37. doi: 10.2478/jagi-2019-0002

Wiehler, L. (2022). *How can AI Regulation be Effectively Enforced?: Comparing Compliance Mechanisms for AI Regulation with a Multiple-criteria Decision Analysis* [Doctoral dissertation]. Fiesole: European University Institute.

Wilkens, U., Langholf, V., Ontrup, G., and Kluge, A. (2021a). "Towards a maturity model of human-centered AI – a reference for AI implementation at the workplace," in *Competence Development and Learning Assistance Systems for the Data-driven Future*, eds W. Sihn, and S. Schlund (Berlin: Gito Verlag), 179–198. doi: 10.30844/wgab_2021_11

Wilkens, U., Reyes, C. C., Treude, T., and Kluge, A. (2021b). *Understandings and Perspectives of Human-centered AI–A Transdisciplinary Literature Review*. Bochum: Frühjahrskongress der Gesellschaft für Arbeitswissenschaf.

Wright, C. (2008). Reinventing human resource management: business partners, internal consultants and the limits to professionalization. *Hum. Relat.* 61, 1063–1086. doi: 10.1177/0018726708094860

Wu, W., Huang, T., and Gong, K. (2020). Ethical principles and governance technology development of AI in China. *Engineering* 6, 302–309. doi: 10.1016/j.eng.2019.12.015

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Commun. ACM* 64, 107–115. doi: 10.1145/3446776

Zhuo, T. Y., Huang, Y., Chen, C., and Xing, Z. (2023). *Red teaming ChatGPT via Jailbreaking: Bias, Robustness, Reliability and Toxicity*. Available online at: https://arxiv.org/abs/2301.12867 (accessed July 16, 2023.).

frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 01 March 2024
DOI 10.3389/frai.2024.1272102

OPEN ACCESS

EDITED BY
Corinna Peifer,
University of Lübeck, Germany

REVIEWED BY
Jairo Hernando Quintero,
Technological Institute of Putumayo, Colombia
Christian Herzog,
University of Lübeck, Germany

*CORRESPONDENCE
Manfred Wannöffel
✉ manfred.wannoeffel@rub.de

RECEIVED 03 August 2023
ACCEPTED 08 February 2024
PUBLISHED 01 March 2024

CITATION
Haipeter T, Wannöffel M, Daus J-T and Schaffarczik S (2024) Human-centered AI through employee participation. *Front. Artif. Intell.* 7:1272102. doi: 10.3389/frai.2024.1272102

Human-centered AI through employee participation

Thomas Haipeter[1], Manfred Wannöffel[2,3]*, Jan-Torge Daus[3] and Sandra Schaffarczik[3]

[1]Faculty of Social Sciences, Institute for Work, Skills and Training, University of Duisburg-Essen, Duisburg, North Rhine-Westphalia, Germany, [2]Department of Social Sciences, Ruhr-University Bochum, Bochum, Germany, [3]Gemeinsame Arbeitsstelle RUB/IGM, Bochum, Germany

This article examines the role of employee participation in AI implementation, focusing on a case study from the German telecommunications sector. Theoretical discussions highlight concepts of employee participation and workplace democracy, emphasizing the normative basis for human-centered AI in Europe. The empirical analysis of the case study demonstrates social practices of human-centered AI and the importance of employee representatives and labor policies in sustainable technology. The contribution is structured into two main parts: first, discussing sociological concepts of employee participation and summarizing the role of works councils in shaping digital technology implementation. Second, focusing on a case study of AI regulations at Deutsche Telekom, highlighting the significant effects of employee participation and co-determination by the group works council in promoting socially sustainable AI implementation which is done via qualitative case analysis. The article highlights the significance of participation and negotiations and gives an example for social partnership relations in AI implementations.

KEYWORDS

human-centered AI, employee participation, works council, ethic rules, company agreement on AI

Introduction

This article is about the role of employee participation in the process of AI implementation both from a theoretical and an empirical point of view by looking at a case study of the telecommunication sector from Germany. On the one hand theoretical outlines give emphasis to some concepts of employee participation and workplace democracy for specifying the normative basis of human-centered AI at work in the European context. On the other hand, the case study analysis presents social practices of human-centered AI to specify criteria of the role of employee representatives and labor policy to implement digital technologies in a sustainable way. In coordinated market economies like the German one management strategies and the implementation of new technologies are strongly shaped by social institutions and regulations of labor relations. However, the way this process of shaping works and the following effects are not determined by the mere existence of social institutions themselves, but by concrete strategies and activities of the actors of labor relations and by the power resources and capabilities these actors can rely on.

The contribution is structured in two steps. Firstly, we will discuss some sociological concepts of employee participation like participation, labor process analysis or production models that can be used for the analysis of employee participation in AI implementation. In this context we will also summarize what is already known about the ways works councils do actively shape the implementation of digital technologies during the last years (Haipeter and Schilling, 2023; Hirsch-Kreinsen, 2023a,b; Kuhlmann, 2023; Pfeiffer, 2023).

Secondly, we will focus on an internationally broadly discussed and recognized practice case of the AI-regulations of "Deutsche Telekom," in which employee participation and the co-determination of the group works council proved to produce rather important effects for a social sustainable implementation of AI, developing three instruments: a Manifesto, a digital roadmap and new form of agile IT company agreements (Bargmann, 2022; Doellgast et al., 2022; Doellgast, 2023; Doellgast and Kämpf, in press). At "Deutsche Telekom," since 2016 the group works council and management have agreed on several company agreements concerning the introduction of digital technologies and especially on the implementation and the use of AI.[1] They have developed an "AI Manifesto" which takes in account the general ethical guidelines of the AI Act of the European Commission and the national AI-Strategy of the German Government.[2] The "AI-Manifesto" intents to structure decision-making processes about the introduction of new AI-systems with new forms of agile company agreements. These new agreements give the works councils an important say in the process of application including a veto-right, and it includes principles that have to be met by new IT-systems. Taking these instruments together, the new forms of agile company agreements and the "AI-Manifesto" represent a particularly far-reaching form of participation of works councils and employees in the telecommunication sector. The following analysis is about potential challenges for employees' participation in the process of the implementation of AI.

The issue of employee participation in the context of AI implementation illuminates the relationship between social institutions and economic practices by focusing on the agency of institutionalized actors. The analysis explores the conditions and activities which allow institutionalized actors to become effective in shaping digital transformations. Effectivity here means both, the fact that the implementation of technologies is influenced by these social actors and that labor policy at company level is an important instrument to protect employment and working conditions. This seems to be even more important as there is an ongoing political debate in the EU and within EU member states about the further development of the EU AI Act and the improvement of legal opportunities for information, consultation and participation.

Key concepts of employee participation

This chapter presents some concepts of employee participation and argues that these concepts have an analytical surplus value for understanding the role and the forms of employee participation might have in the implementation of AI or other forms of digital technologies. The first of these concepts is *participation* itself, which is traditionally among the key concepts of labor and industrial sociology. In former times, participation has been regarded as a quality of collective action of trade unions or other representations of employee interests. In this sense it was regarded as part of "industrial citizenship rights" of employees (Marshall, 1950). In any case, in this view collective action took place beyond the limits of the individual organization of companies: either like in the British tradition of "industrial democracy," as a quality of collective bargaining between employers and—independent as well as professionalized—trade unions; or like in the German tradition of "economic democracy," as a result of trade union participation in the centralized planning of the economy. There was no independent role of direct participation on company or establishment level given in these overarching concepts (Haipeter, 2019a).

However, and on the contrary to this, nowadays participation is recognized as an independent element of labor relations *besides* collective bargaining. Whereas the latter is about collective contracting of labor standards, the former is about having a say in the concrete conditions under which labor is used in the organization of the labor process. This means that direct participation is the cornerstone of what can be called "democracy at work," based on certain status rights workers can dispose of beyond the contractual conditions of the sale of labor power, be they legally and/or collectively agreed (Dukes and Streeck, 2023).

Participation as an analytical concept of its own emerged during the 1960s, driven both by the fact that in several European countries statutory rights of participation on establishment and company level have been implemented in course of the postwar reconstruction of the economies and driven by the critique of the representative structures of the labor movements that developed during the 1960s. From then on, participation has been regarded as a democratic element within the economy that is based on influencing firms' decision-making both in a representative way by labor representatives and in a direct way by employees themselves within establishments and companies. As such, participation has become an important concept in comparative research about industrial relations and at the same time an interdisciplinary concept also used in organization or HR theory (Wilkinson, 2011).

Moreover, participation in this sense can rest on very different forms, ranging from information to consultation and to codetermination. In the case of information, workers or their representatives have to be informed about managerial decisions; consultation means that they are able to articulate their interest about these decisions which can then be included or ignored in the decision-making process; and in the case of codetermination, finally, the decision cannot be made without the consent of the workers. In most of the European countries with statutory participation rights, these rights refer only to information and consultation (Haipeter, 2019a). This is also true for the European level and its core institution of European Works Councils. One of the most important exceptions from this rule is Germany, where the statutory rights of participation also include codetermination, at least with respect to certain topics of the implementation of new technologies.

The German case is instructive for our analysis, both because it includes the most developed forms of participation in the sense

1 In the multi-level system of co-determination (local works council, group works council, European works council), the group works council at Telekom group level leads the negotiations with management on the introduction of AI.

2 See: https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence?, https://www.ki-strategie-deutschland.de/home.html (06-26-2023).

of codetermination and because we refer to it in our case study of the role of the group works council of "Deutsche Telekom." Codetermination rights, as they are listed mainly in the renowned § 87 of the German Works Constitutions Act, extend over several issues, ranging from the distribution and position of working times working times to wage methods or the organization of teamwork. Of special importance for digitalization issues are the § 87.6, which ensures codetermination in case of the introduction and application of technical instruments that might be used to monitor the conduct and performance of employees, and the § 87.14, which is about codetermination on the regulation of mobile work which is based on IT-technologies. Furthermore, the § 80.3 which was adopted in the Works Council Modernization Act of 2021 nowadays gives the works councils the opportunity to consult an external expert in case AI systems are introduced and have to be assessed by the works councils without any permission by the management.

Participation in this sense can be regarded as a bundle of institutionalized collective status rights of employees. However, as legal norms these rights tell us little about how labor policy does function in concrete social situations and in how far they actually shape social practices. Here a second important concept to deal with employee participation comes into play, the notion of the *labor process* as a terrain of politics, conflicts and contests. This view has been developed in the British Labor Process Debate, which stresses the aspect of control in managerial strategies as a means to regulate and monitor the labor process and to cope with the problem of transforming labor power into concrete labor (Thompson, 1990).

However, as this research tradition has shown rather different strategies of control can be distinguished. Control does not mean that management tries to monitor every aspect of the labor process. Instead, control strategies may range between the extreme poles of direct control—like in Taylorist forms of scientific management with high levels of division of work, rigid separation of execution and control and the concentration of the knowledge of the labor process in the hands of management—on the one and responsible autonomy on the other hand, the latter giving the workers broad leeway to apply their qualifications and knowledge (Friedmann, 1977). In this perspective, management not only has choices to make, but there is also room for contestations, negotiations, and compromises between management and labor about control issues at work.

In a complementary way, Edwards (1979) has distinguished three forms of control as an expression of the "structured antagonism" that characterizes the relationship between labor and capital on the shopfloor which is constantly negotiated and re-negotiated. Control in this sense means a system of political regulation. The three forms of control systems according to Edwards are, first simple or personal control by managers and superiors supervising the labor process, technical control by the demands and connections of technological artifacts and machines like the assembly line, and, finally, bureaucratic control by the institutionalization of control in the form of job descriptions or rules of promotion. The two latter forms depersonalize and, in this way, according to Edwards, mystify the control relationships as independent technological necessities or institutional rules.

This analysis connects control and conflicts about control issues with the aspect of consent as a precondition for stable control systems. According to Burawoy (1985), work contexts are characterized by three strongly connected dimensions: the economic dimension of the production of things or services, the political dimension in the sense of the production of social relations, and the ideological dimension by producing experiences of those relations. Interests of workers and management are coordinated within the political and ideological dimensions of work on the shop floor, producing a hegemony within more or less stable work regimes which are not continually contested. This mostly applies to regimes based on a more or less stable balance of power between labor and capital, much less however to coercive or despotic hegemonial regimes in which power and coercion are visible and may lead to contention.

Given these understandings of the labor process, labor process theory suggests to analyse the digitalization of the labor process with respect to issues like the skilling or deskilling of labor, the effects on the autonomy and responsibility of the workers, the control regimes and the ways digital technologies contribute to or modify existing control strategies and, finally, to the production of consent about the implementation of digital technologies in the labor process (also Briken et al., 2017).

However, as Thompson and Laaser (2021) argue, looking at technology it makes sense to distinguish first and second order strategies of management, with first order strategies concerning the development and adoption of technological systems in interactions between firms, state actors and scientific-professional domains, whereas second order strategies are about the implementation of technologies and concrete strategies of control and about negotiations and contestations of these strategies. Furthermore, in line with the concept of *production models* which connects the dimensions of company strategies including finance and product strategies, process organization including the labor process, and labor relations between management and labor representatives (Boyer and Freyssenet, 2003), the authors argue that the control regime is embedded in a regulatory regime of labor regulation and an accumulation regime including conditions of competition and finance.

As research has shown, employee participation and the way it can be implemented in conflicts about autonomy, control or qualifications largely depends on the *power resources* the actors can rely on in the labor process (Schmalz and Dörre, 2014). The most important of these resources for an effective employee participation are: (1) structural power, which is based on market and organizational positions of employees and which gives them either individual power or power for the collective actors in which they are organized; (2) organizational power in terms of high trade unions density or the ability to mobilize workers in concrete conflicts related to issues of participation; and (3), institutional power, which is based on the legal rights of employee representation in companies, both in terms of the organization of these actors and the concrete rights of information, consultation or more advanced forms of participation they can dispose of. It has been stressed in literature that in the context of digitalization a fourth form of power may play an important role, which is discursive power which shapes the way digital technologies are interpreted, either as

instruments of autonomy and improvements of working conditions or as instruments of competitiveness, rationalization and control (Kuhlmann and Rüb, 2020).

However, this analysis is about potential topics and issues without looking at the concrete agency of labor representatives and workers and the conflicts, negotiations or new forms of consent that might develop around these issues. This kind of analysis needs in-depth case studies also in combination with industry studies in order to better understand business policies on digitalization and the role of labor relations and regulations the implementation is embedded in. This is what this article tries to show for the case of the role of the group works councils of the Deutsche Telekom. Before we do this, we will give a short overview on the findings concerning digitalization and the role of codetermination and trade unions in Germany.

Works councils and employee participation in digitalisation processes

What do we know about the role of works councils and trade unions in German play in digitalization processes? Do they participate actively in these processes, do they shape conflicts and consent in the labor process, and do they influence the development of production models? The findings on this question are quite ambiguous at first glance (Kuhlmann, 2023). This is especially true with regard to works councils, which as codetermination actors are at the center of participation in negotiations on digitalization in the labor process (Haipeter and Schilling, 2023). First of all, it can be generally stated that codetermination represents a "regulatory environment" for the implementation of digital technologies, insofar as the negotiations between the collective actors of management and employee representatives in the company enable certain forms of use of the technologies and limit others (Krzywdzinski et al., 2022). This has been empirically demonstrated not least with respect to wearables and digital assistance systems, which were introduced in the logistic sector and in the production areas of the manufacturing sector. In this case, works councils have proven to be able to negotiate restrictions on data-based performance control, based on their codetermination rights and accepting rationalization effects as the baseline of compromise (Falkenberg, 2021; Krzywdzinski et al., 2022).

However, this finding still says little about concrete strategies and choices works councils have developed in dealing with digitalization. Most studies find that works councils deal with digitization projects in a mostly reactive manner. Reactive means that works councils primarily develop protection claims and try to reduce or compensate for the negative consequences digitalization may have for employment and working conditions. These patterns of action can be distinguished from more active attempts to gain influence on the design of technology and the associated work organization, an approach that seems to be pursued much less frequently.

As Kuhlmann and Voskamp (2019) show in their study on digitalization in mechanical engineering, company representatives tend to be unsettled and overwhelmed, especially in SMEs, due to a lack of resources, limited technical competences and a lack of involvement by management. The situation may be different in larger companies where resources are better and management is more cooperative. Accordingly, the authors contrast strategies of works councils with the attitude of waiting and retreating to consolidated positions of action on the one hand, and claims of proactive participation on the other hand, in the context of which the attempt is made to exert influence on projects about work and organization.

In their study on conflicts over digitization in companies, Rüb et al. (2021) emphasize that the claim of actively influencing digitization processes and developing one's own strategic claims can at best be pursued by resource-rich works councils in large companies, while in smaller companies' resource bottlenecks of the works councils with regard to time, personnel, knowledge and assertiveness make it difficult to help shape the change. Therefore, a reactive protection policy remains a central and for many works councils the only strategy for dealing with digitalization, especially as a competitive discourse dominates in many companies and is also accepted by the works councils, which classifies digital technologies and the associated rationalization and productivity potentials as an unavoidable precondition for competitiveness as well as maintaining locations and employment.

This assessment is shared by Bahnmüller et al. (2023) in their recent analysis of digitization-intensive companies in the metal industry. The authors note that works council action in digitalization processes is generally reactive and aimed at monitoring. Active support for digitization projects is just as uncommon as participation of works councils in teams which are planning and developing digitalization. On the one hand, this is due to resource bottlenecks of the works councils, which do not allow for more extensive activities, but on the other hand also to the assessment that in this way the employees' interests can be represented quite effectively, especially by negotiating employment effects and performance controls.

These findings are in line with the results of the survey conducted as part of the IG Metall—the German metalworkers' union "transformation atlas" (Gerst, 2020). According to this survey, only a smaller proportion of works councils is informed about and involved in change projects at an early stage. From a trade union perspective, Gerst assesses this mode of interest representation by the works councils as "disastrous," because from his point of view only through more active involvement can employment security and good working conditions be influenced in the longer term in the interests of the employees.

However, there are examples of works councils taking a more active role in shaping digitalization. According to Rego (2022), the prerequisites for this are both a high strategic importance of digitalization as a field of action for the works councils and a strong resource position of the works councils. Under these conditions, the works councils can develop a more active stance on digitalization and develop strategies and claims against company management. There are two conditions in particular that are considered important for works councils to strategically shape digitalization: on the one hand to organize their own work effectively based on clustering competencies in thematic committees, on the other hand to organize direct employee participation within representative works councils' codetermination as a resource

for mobilizing the competencies of the workers as experts of their work.

These findings are in line with the analysis of capabilities by Lévesque and Murray (2010), who stress two aspects of capabilities to participate by trade unions or works councils. The first aspect is the internal reorganization of works council work by restructuring bodies and committees, setting up project and working groups, ensuring the internal knowledge acquisition of workers' representatives through training, bringing in external expertise through specialists or also by strategically planning the composition of the works council body from the different specialist areas of a company (see also Niewerth and Massolle, 2022). The second and complementary aspect is the participation of employees. This is basically about using their expert knowledge and at the same time increasing the legitimacy of the representation of interests (Bella et al., 2022; Niewerth and Massolle, 2022). However, these practices seem to be little practiced beyond the boundaries of particularly active works councils' committees. According to Bahnmüller et al. (2023), works councils support forms of management participation within the framework of lean concepts, but do not practice employee involvement as a systematic element of their own work.

In addition, there is a third aspect that is rarely considered in the study of interest representation which is important in the German case, the division of labor between works councils and trade unions as an important basis for the ability of works councils to act. This division of labor is traditionally characterized by mutual support services: trade unions qualify works councils, help them with specific requests, lend them organizational power and relieve them by concluding collective agreements, while conversely works councils monitor compliance with collective agreements, regulate company- and workplace-related issues and recruit members for the trade unions. In this pattern of division of labor, the competences of the trade unions were only called upon by the works councils when needed, an approach that, according to trade union assessments, is no longer sustainable and should be replaced by a more active positioning of the trade unions in order to create the basis for a broader claim of the trade unions to shape the future (Gerst, 2020).

German trade unions have focused these activities in projects in which they try to strengthen the capabilities of the works councils to play a more active role in negotiations and to develop strategies of their own as alternatives to management strategies. In this context several projects have been implemented by the metalworkers' trade union IG Metall which have tried to enable works councils and especially those works councils from SMEs with little resources and capabilities to participate in digitalization issues more actively, to negotiate agreements on how to deal with digitalization projects and to develop own concepts of business strategies based on digital technologies, the most important of them the project "Arbeit und Innovation" together with the Learning Factory of the Chair of Production Systems of Ruhr-University of Bochum (work and innovation) and "Arbeit 2020" (*work 2020*). In the latter project, works councils have been supported by external consultants and trade union officials by up to ten workshops in each case which took place on establishment levels (Haipeter, 2019b). These workshops tried to realize three different goals: Firstly, to develop a digitalization map of the establishment together with employees, secondly, to discuss the political implications of these findings and to identify core topics like employment protection, problems of qualification, deteriorations of working conditions or management problems; and, thirdly, to negotiate these issues with management, trying to pave the way for an agreement which strengthens the opportunities of works councils to participate in digitalization projects and to bring in their own concepts and social aspects. In total, nearly 100 companies and works councils attended in the project, and around 20 agreements have been concluded between works councils and management which focused mainly on procedural rights for the works councils to participate in digitalization projects.

Projects like "work 2020" show that trade unions can give important stimuli to activate works councils mainly from smaller companies to develop new competencies and capabilities to deal more strategically with digitalization issues and to develop a more active approaches of participation. Moreover, they show that negotiated participation in the case of digitalization is less about substantial norms and more about procedural rights of works councils and about opportunities to attend and influence processes of innovation. At the same time, this means that participation in digitalization issues, if it takes place at all, is a continuous task for works councils which requires capabilities of their own in terms of reorganizing the work within works councils' committees (see also Rego et al., 2021). This is a core precondition in terms of agency for the institution of codetermination to shape digitalization and the introduction of AI in firms.

AI implementation—challenges for works councils

Against the background of the theoretical key concepts and actual empirical findings on the importance of labor relations and especially of labor politics of workers' representatives in the context of the introduction of digital technologies, the following chapter gives a deeper insight how works councils deal or are able to deal with the introduction of AI. Besides the shift to remote working in the wake of the COVID-19 pandemic (Kötter et al., 2023), the introduction of AI is certainly one of the biggest current challenges for works councils, as AI systems could lead to substantial changes in work processes and new qualification requirements for employees. In general, the implementation of AI can have a direct impact on employees and their activities (human-centered) or primarily on technical processes and thus only secondarily on employees (technology-centered) (Huchler, 2023; Pfeiffer, 2023). Taking these possible different paths in account, the introduction of AI in the company confronts works councils with vital challenges. The first question that arises are the competences necessary for understanding and dealing with AI, as well as anticipating the far-reaching changes that the introduction of AI can mean for work processes.

A clear stance is needed that pushes for the enforcement of co-determination rights regarding AI. Often employee representatives are overwhelmed in the first step and realize that there is no suitable set of rules for such a case. On the part of the employees, the

committees might be confronted with reservations and fears, even though there are not yet reliable figures on the long-term effects of AI on employment (Ver.di, 2020). A comprehensive stakeholder sensitization is needed, which includes in particular a technology impact assessment.

In addition, as in the case of the AI implementation at the German company Siemens which will be analyzed later in this article, AI applications are often not readily recognizable and are mixed with automation and general digitalization processes (Grasy and Seibold, 2023). The fact that a generally applicable and comprehensive definition is often still lacking (Höfers and Schröder, 2022) rises points of conflict where the employer side restricted the concept of AI to self-learning systems alone and thus wanted to undermine the right of the works council to have a say (Grasy and Seibold, 2023). However, since the amendment Works Council Constitution Act in 2021 (BetrVG 80, 3) allows the co-determination body to call in experts to advise it.

Taking this legal base of AI implementation in account, international comparative studies underline, that in German cases of AI implementation is a tendency toward social partnership solutions, which often take a similar path (Doellgast and Kämpf, in press). AI is often seen as a "cross-cutting issue" with effects on areas of employment and labor conditions as well as collective bargaining policy. A particular argument here is the reference to the EU AI Act, which also addresses the ethical basis of "AI made in the EU" and excludes certain types of AI (high risk) from the outset. At the same time, AI systems are often still a "black box"—whether personal data can be collected, for example, can often only be examined after purchase (Grasy and Seibold, 2023). Following theses authors, co-determination must become a direct part of the introduction process.

At the same time, the introduction of AI can offer an opportunity to enter into negotiations, e.g., to force further training and retraining, but it can also lead to more stress and anxiety (Doellgast, 2022). Trade unions (and in the German case, first of all works councils) are confronted with three main problems: the threat of job losses, special requirements for data protection and the challenge of organizing outsourced employees, e.g., in subcontractors. Europe and Germany have comparatively strong regulations with regard to data protection. Solutions to these problems can be attempts to influence government legislation; negotiating new labor standards through trade unions; and at plant level company agreements. This level and the challenges and approaches associated with it will be examined further in the following example of Deutsche Telekom.

The qualitative case analysis—process of developing a works agreement for artificial intelligence systems

In the context of a qualitative analysis the question is explored of how works councils can have a say before the introduction of new AI systems already begins. Since valid works agreements on IT are no longer sufficient when AI is already introduced, the core criteria for a model company agreement on AI are being worked out during this analysis. In order to understand the contextual conditions and the participation of works councils in the introduction of AI solutions at Deutsche Telekom, a comprehensive document analysis of company agreements and open guideline interviews with members of the group works council (GWC) were used. The aim was to draw on the experiential knowledge of workers' representatives to enable a reconstruction of the decision-making process. Deutsche Telekom was also a project partner in the BMAS-funded project "humAIn work.lab," which investigated risks and opportunities in the application of AI at work (in the period from 2020–2023). The underlying transfer research concept enables the work-oriented implementation of research projects with a focus on the transfer of knowledge between scientific disciplines and practitioners. This knowledge transfer as a constitutive component of the research process contributes significantly to an interlocking of research and social practice (Schäfer et al., 2022, p. 129–132). In general, this method provides "exclusive insights into the complexity of structural contexts and processes of change in systems of action, such as decision-making structures and problem-solving in organizations and institutions" (Liebold and Trinczek, 2009, p. 53). To be able to track the work steps of a works council committee in this context, a works council committee was to be accompanied at intervals of several weeks over a period of 2 years. In the course of intensive cooperation (Schäfer et al., 2022), with the group works council of Deutsche Telekom Service GmbH, the data collected in advance was condensed during the field analysis through the perspective of active works councils. The dialogic interviews with works council members of Deutsche Telekom Service GmbH are recorded in detailed protocols and supplementary visual material and evaluated in several phases. The analysis of the collected data is aimed at identifying core criteria that facilitate the development of company agreements on AI.

Case study Deutsche Telekom

Operating agreement for artificial intelligence systems

The strategy of Deutsche Telekom's group works council (GWC) was chosen as a case study for two reasons: because this company develops Artificial Intelligence (AI) based tools which makes it a vanguard company of the German IT sectors and, on the other hand, because its works council plays a very active role in the regulating and shaping the AI introduction processes within the company (Doellgast and Kämpf, in press).

The company offers products and services in the areas of fixed network, mobile telephony, Internet and Internet TV for private customers as well as information and communication technology solutions for major and business customers. The former public company Deutsche Telekom was privatized in 1996, and in 2022 considered the largest telecommunications company in Europe. The German government still holds nearly 32% of the company's stock in 2022 and counts with 220,000 employees worldwide and more than 90,000 of them in German locations. Because of its history as a former public company, the Deutsche Telekom AG is still highly unionized by nearly 80%, despite being a high-tech company.

Codetermination at the Deutsche Telekom takes place in the form of a multilevel system, composed of local works councils, central works councils for the divisions and subsidiaries of the company and the group works council, which is composed of members of the different central works councils. Issues related to the implementation of IT systems are dealt with in the group works councils as many IT systems are used in the whole group and not only in certain divisions or subsidiaries of the corporation. In total, the group works council (GWC) consists of 27 members from 10 delegate areas. The GWC has established a special committee dealing with IT issues, the IT committee, which is composed of 4 GWC members and other works councils from the central works councils and from local works councils which are at the same time experts in dealing with IT issues (Bargmann, 2022).[3]

A core approach toward AI developed in the GWC of Deutsche Telekom is that AI is not to be regarded as a finished technology, but as a learning system of information technology. In this view, AI evolves to perform tasks, optimizes itself and solves problems by independently recognizing patterns, drawing conclusions and preparing or making decisions. Taking these patterns into account, the introduction and the use of AI is an ongoing process of a deep technological transformation.

The AI Manifesto

In this context, in October 2022 the so called "AI Manifesto" was concluded between the Deutsche Telekom management and the group works council. The "AI Manifesto" is an agreement between the GWC and the company management about the introduction and implementation of AI within all the section of the company. Apart from regulating AI implementation, the Manifesto at the same time can be regarded as a new type of agreement between works councils and management because it represents a new forms of agile company agreement.[4]

Basically, the agreement refers to national and international (EU) legal regulations and (technical) standards and supplements the Group Works Agreement on IT Systems, the Group's Digital Ethics Guidelines for dealing with artificial intelligence and General Data Protection Regulation (DGPR). Basic positions were laid down also referring to the latest legal amendment of the German Work Constitution Act from July 2021 that stipulates that employees have to be informed about possible interactions with learning machines, that personnel-relevant decisions must not be made by AI or that AI is not allowed to be used for surveillance. Based on this, common goals and procedures were agreed upon concerning on the introduction and use of artificial intelligence the generally applicable regulatory framework, quality requirements, dealing with risks or the introduction of a group of experts composed of management and works councils. Another important point of the agreement is that it includes the rule that employees of Deutsche Telekom have to be at the center of all operational decision-making process concerning AI. In detail, the main principles of the Manifesto are the following:

First, the interaction between employees and learning machines has to be designed in such a way that employees are informed about the fact that they are interacting with such a machine. In line with the already existing agreements on IT Systems, the Manifesto says that employees have to be protected against machine control of performance and behavior and prohibit the use of unauthorized humane data. Only human decision makers are attributed the right to draw conclusions relevant to human resources that could have legal effects on employees or significantly influence them in a similar way. Employees who are indirectly affected by machine conclusions with personal effects can request a review of the system decision from those responsible. Furthermore, according to the agreement AI systems will not be used to analyze, influence or control employees' emotions or mental state. Employee biometric data and AI systems designed to improve employee wellbeing will only be used if permitted by other company agreements.

Besides these more basic rules, the *AI Manifesto* includes procedural rules about how to cope with the implementation of AI systems.

At first management and the group works council agreed on quality, trust factors, and quality checks of AI in which also works council members are involved: Legal and regulatory compliance of AI solutions, transparency, compatibility with the Digital Ethics "AI Guidelines," usefulness in the performance process, risk appropriateness, controllability, protection of personal rights, ergonomics, social compatibility, good work, robustness, and sustainability.

Secondly, the agreements stipulate that the group works councils should be informed in early stage about the data sources of the AI system and assessments of the informative value and integrity of the data system description, model of the AI system, plans for evaluating the model quality in ongoing operation and emergency concept, depending on the respective risk classification and planning phase. Works councils can demand unscheduled monitoring from those responsible for the system if there are indications that the system is not being used in accordance with this agreement.

Thirdly, the agreement states that a joint AI-expert group with the management (4 members) has to be implemented. This group receives, together with the GWC, the information on the results and methods of system training and testing. The expert group is continuously involved in the development of impact assessment procedures, standards for the assessment of risk dimensions and their probabilities of occurrence. Apart from this and depending on special issues, group works council is allowed to call in further experts according to § 80.3 Work Constitution Act to create generalizable procedures for co-determination and quality assurance on AI systems. Finally, the operational functional managers for the AI applications and representatives of the works councils (IT Committee) will be continuously qualified to put into practice this Manifesto (Höfers and Schröder, 2022).[5]

3 See; Deutsche Telekom (2023): HR Factbook 2022, Menschen. Fakten. Entwicklungen.

4 See: https://www.telekom.com/de/konzern/details/telekom-verpflichtet-sich-auf-ki-ethik-1025794 (15. 10. 2023).

5 Source: Manifesto between the Deutsche Telekom Group and the Group Works Council on the introduction and use of information technology systems, October 21, 2022.

The AI Manifesto in practice

In this context the group works council has developed a pyramid of criticality levels of AI based applications and systems, which refers mainly to the AI-Strategy of the current Federal Government and determines the damage potential of an AI and provides for measures and actions accordingly (see Figure 1).[6] Depending on the risk classification of planned AI applications, different possible actions for the expert group are defined. The potential for harm of the application is assessed in five levels. Level 1 (green) refers to the introduction of AI with no or little potential for harm to employees in the sense that it does not interfere with personal basic rights. In this case, no separate regulatory measures in company agreements are required on the part of the works council. Level 2 (yellow) describes a certain potential for harm by reducing the decision-making autonomy of the employee through digital twins. In this case, management is obliged to comply with certain transparency obligations and to carry out a risk impact assessment. This includes specific control and evaluation procedures. Levels 3 and 4 (orange) indicate AI applications with regular and significant potential for harm by the potential use of sensor technology that detects and processes employee behavior. These are either reviewed through ex-ante approval procedures or prohibited if necessary. Level 5 (red) indicates an area of AI application that is considered unacceptable and that is rejected by the works council. These AI applications would have the potential to monitor the employee's behavior or performance with corresponding consequences for pay development (which is forbidden par § 87.6 BetrVG). If these technical possibilities can be ruled out through an evaluation, the group works council can partially agree to this AI application afterwards. In essence, level 5 covers with all the regulatory areas of section 87 (1) 6 and 10 of the Works Constitution Act, which are subject to the co-determination of the works council. In this case management is not allowed to introduce this AI without its consent.

The criticality levels marked in Figure 1 are the first step to an operationalization of the programmatic statements in the Manifesto. In this way, the management of Telekom and the GWC have developed an ethical framework that will enable them to introduce AI systems in a dialogue-based and structured manner (Höfers and Schröder, 2022).

The second step of operationalization of the AI-Manifesto and the pyramid of critical levels was the development of a so called "digital roadmap" (Doellgast and Kämpf, in press). This roadmap defines steps of participation the GWC can potentially make use of, in line with the review of the rating of the AI. These steps of participation are about renegotiating existing agreements on IT systems based on the results of the assessments made; given this, the digital roadmap can be regarded as "learning" regulation which allows to adapt regulations to new facts. In practice this means that after the initial information by the management has taken place, the GWC participates in the development of so-called system profiles, which form the basis for both a review of the system and a possible need for action by the works council. If, after documenting the audit results, it is determined that there is no need for action (usually at criticality level 1), the profile is closed and the existing IT company agreement should not be renegotiated. If, however, a need for action is identified after the audit (usually at criticality level 2), elements of the IT company agreement have to be renegotiated (see Figure 2).

Therefore, the digital roadmap presents the base for third step of AI introduction by Deutsche Telekom, the development of new forms of agile company agreements on IT systems. These agile agreements can be regarded as a strategic change toward a digitalisation of co-determination processes. At its core is a profile procedure for IT systems. Linked to the Manifesto programmatic, it is controlled by a project management software JIRA@BR (Bargmann, 2022) which is based on a new version of the GWC agreement on the planning, introduction, use and modification of IT Systems (GWCA IT Systems) from March 2021 and on the GWC on Digital Cooperation (GWCA DC).

While so far, the GWC used to prepare separate, specific company agreements for each new digital tool, now on the base of the digital roadmap the works councils are able to develop new and comprehensive company agreements that sets labor standards. In this context the GWC members have recognized that the preparation of independent from each other and isolated company agreements on continuous technological innovation is too time consuming, especially in view of the rapid development of AI. New agile company agreements include individual rules which always apply, while other sections are to be understood as core principles which should always be taken into account in the context of technical innovation processes. This refers mainly to the content of § 87 Works Constitution Act and the protection of the basic personal rights of employees. In this context, the rights of co-determination of the works councils are no longer contested in negotiations with the management; they are taken as given by the procedural rules.

At the same time, these forms of accelerated co-determination procedures offer advantages for management, as it allows finally to speed up the introduction of digital technologies in general and AI in concrete terms. When new AI systems are introduced, the following process of labor policy applies: First, initial information of the GWC by the management at the earliest possible opportunity; second, draw up a profile of the program; and third, check need for action referring to the question if the basic rights of the employees are met. The AI implementation is thereby examined by the GWC from an application perspective.

GWC members reported, that veto rights until today have rarely to be used—often rather in the case of misunderstandings. However active control of the process is still important in the opinion of the workers' representatives. The result is finally an agile "dual model" of IT co-determination. It is characterized firstly by a general, fundamental and overarching set of rules applicable to all IT systems (GCA IT systems), which are and no longer negotiated, and secondly the concentration in the day-to-day business of ongoing co-determination on those IT systems that require deviations from these core principles. The procedural core of this is the so-called system profile:

Figure 3 illustrates this agile model of co-determination concerning the introduction of IT-systems established by the AI Manifesto. The works council has to be involved from the beginning of the introduction process. An important role is played by the technology assessment of AI (system profile). Possible rationalization processes resulting from the use of AI are also dealt

6 See: https://www.ki-strategie-deutschland.de/home.html.

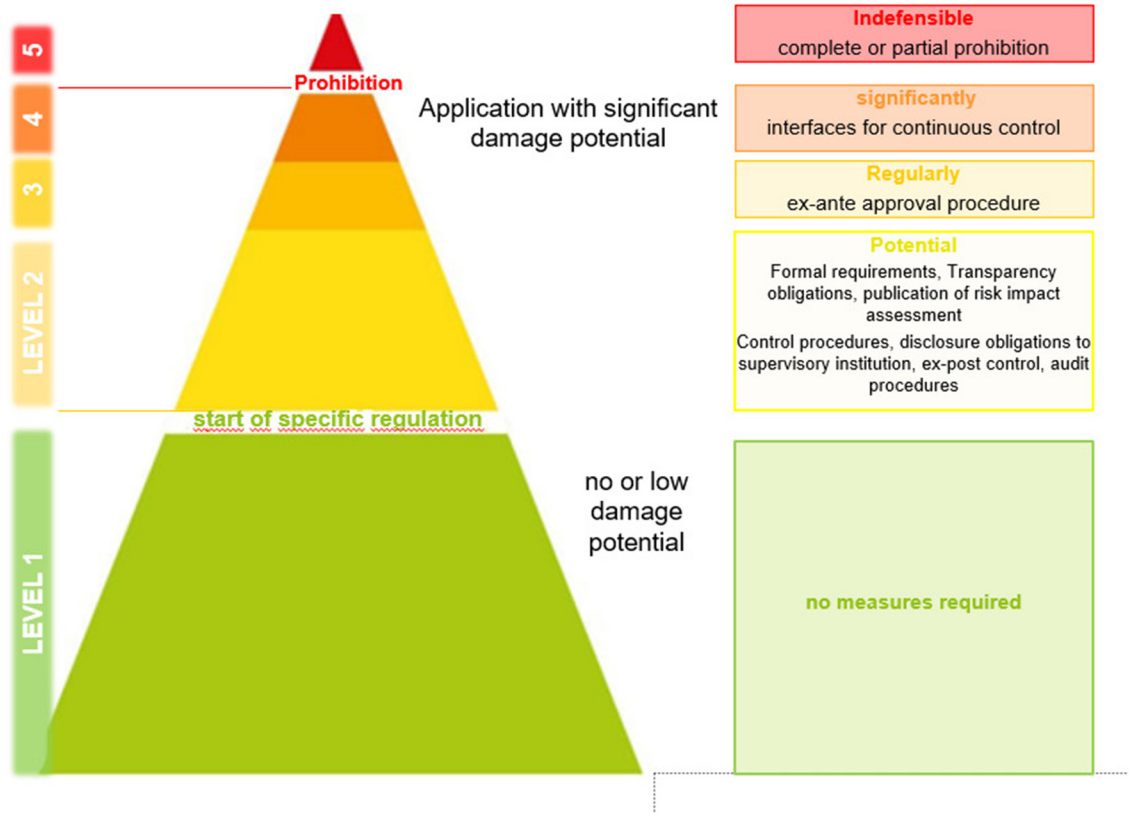

FIGURE 1
Criticality levels of AI introduction. Source: Deutsche Telekom.

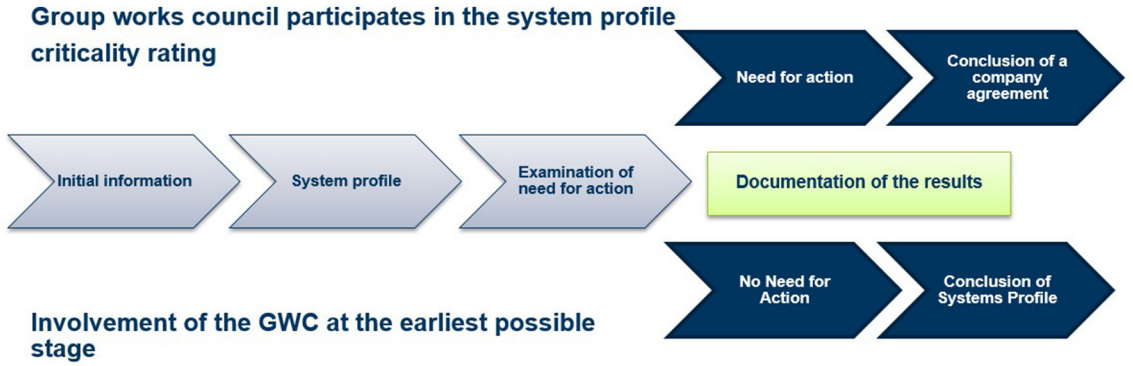

FIGURE 2
Digital roadmap. Source: Deutsche Telekom.

with proactively, because management has to present the (planned) digitalization goals in advance. The works council is informed in this regard and then can become active itself. This is particularly important as the system automatically assigns enough work so that a possible reduction in the workload of individual employees cannot be identified easily. Moreover, sometimes it is not even clear which tasks are omitted or have already been taken over by AI. In order to cope with these sophisticated problems, works council members receive continuous trainings. The costs of these are fully covered by the employer under section 37.6 of the German Works Constitution Act. There is also a regular exchange with the employee representatives on the supervisory board (Höfers and Schröder, 2022).

These three steps of participation in the context of AI-introduction underline a strategic re-orientation of co-determination on AI issues in the Deutsche Telekom AG. It presents new and innovative approaches of agile procedural rules for co-determination of works councils. However, implementation

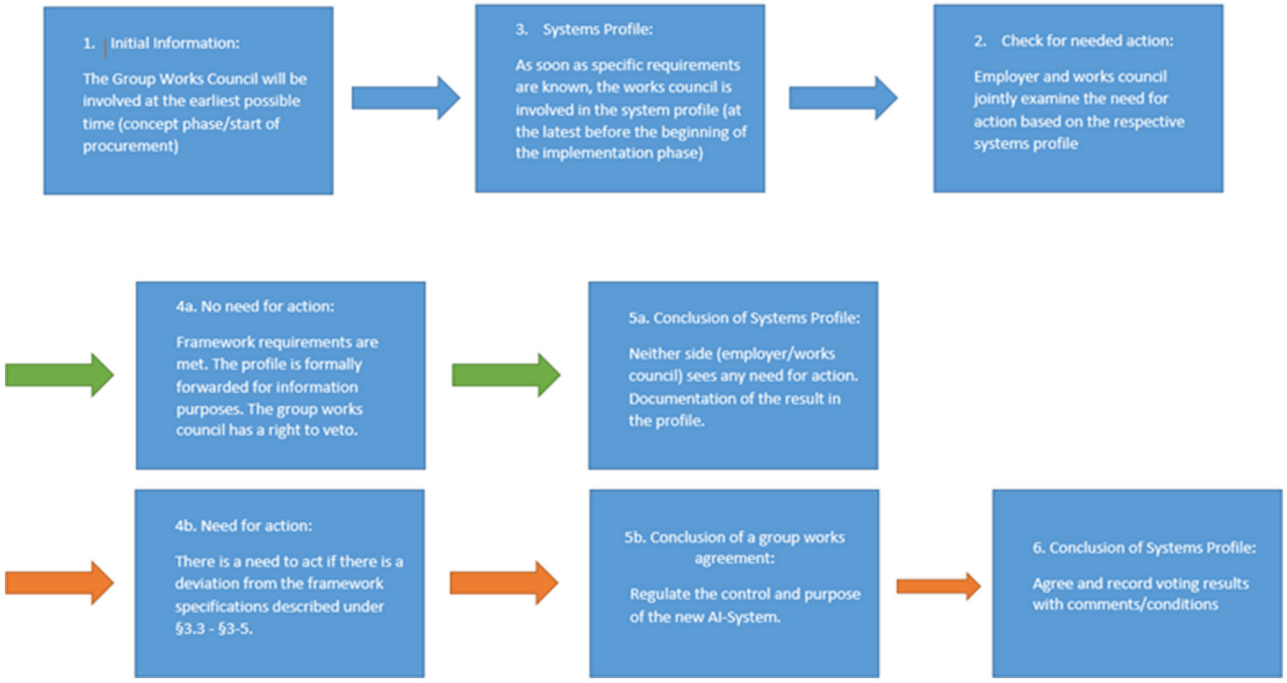

FIGURE 3
Operational practice: IT co-determination process including AI. Source: Deutsche Telekom, own preparation.

of the new regulations 1 year after the conclusion of the Manifesto are still in a learning phase and there is still a need for further empirical analysis of how these agreements work in practice.

Contextualization of the case study

The qualitative analysis has described the way in which the group works council at Telekom has reached a social partnership agreement in the context of AI implementation. But companies of the telecommunication sector play a special role here (Doellgast, 2022), so the following chapter refers to two other actual examples with their approaches to the introduction of AI to finally contextualize the Telekom case.

The first example is Siemens, the largest industrial manufacturing company in Europe, specialized in industrial automation and industrial software. Siemens is already developing and using AI itself (e.g., in personnel processing; as a supporting and relieving chatbot), but in comparison with Deutsche Telekom still has no fundamental company agreement on AI with regard to co-determination (Grasy and Seibold, 2023). Until 2023 only preliminary work has been done by general works council. All relevant functions and forms of use of the respective AI applications have to be presented in profiles, so called "AI cards." The applications and its tasks as well as the possible consequences on employment and labor conditions are to be made comprehensible and clear in this way and help to reduce uncertainty among the employees.

Although the effects of new applications on employees and the resulting measures can be grasped in this way, the general works council is only acting after the introduction of AI reactively until today. At the same time, the group works council has been able to establish guidelines about data protection and data storage and, more broadly, basic ethical considerations that are recognized by the Siemens management. For definitional standards, however, the committee is placing more expectations in definitions from the EU. The group works council was accompanied in this process by an expert team of the German metal union, IG Metall. At this stage of development, Siemens is still relying on "weak" AI, i.e., rather AI assistance, which in forms of chatbots is so far only intended to relieve employees internally. Nevertheless, this could also be a step toward job cuts, as the first contact with customers could also be taken over by a bot (Doellgast, 2022).

The second example is International Business Machines Corporation (IBM), Germany. This world leading company of IT-services is already one step ahead of Siemens. IBM has reached a company agreement on AI since 2020. Similar to the case of Deutsche Telekom, they group works council and the management have developed a framework agreement on the conditions for the introduction and operation of IT systems, which explicitly excluded work performance and behavioral control of the employees. This company agreement later became the basis for the group agreement especially on AI tools—a process which, according to Doellgast et al. (2022), is relatively known in the German system of labor relations. At the same time, internal ethics guidelines also existed in advance. The EU Ethical Principles for Trustworthy AI and the study by the

Bundestag's Enquete Commission on the Potential of AI were also consulted.

Like in the Telekom case, also at the IBM group works council there was ultimately great interest to reach a company agreement, which took place in an open and solution-oriented process with the management. Representatives of the group works council and the representatives of the severely disabled employees, together with HR staff and in-house IT specialists of IBM, were able to learn about the technical basis of AI in a joint series of workshops and at the same time collect topics for a possible company agreement. The focus was on the primacy of human decision-making and the possibilities for intervention as well as exclusion of social discrimination. Like in the case of Deutsche Telekom, damage categories or risk clusters were established here, which demonstrates a certain way of dealing with the respective application of AI. Representatives of IBM's works council have announced that it has been one of the first large companies in Germany which has established a company agreement on AI. Analogous to Siemens, IBM also works with "AI fact sheets" and has similar to Deutsche Telekom—an ethics council that examines new AI applications. Nevertheless, the definition of AI and the question of when it is an intelligent system has not yet been comprehensively clarified in the case of IBM (Remers, 2023).

The examples of Siemens and IBM Germany also underline some general results of the qualitative analysis on the Telekom case. Ethic frameworks like the AI- Manifesto and instruments and methods like the digital roadmap seem to be able to support the development of new types of agile company agreements in the context the introduction of AI solutions. Looking at the broader landscape of German labor relations and codetermination and the opportunities to learn from the examples of these large companies, it should be reflected that the power resources of workers' representatives to exert influence in the development of AI-projects in these companies are much greater than they are in the procurement of AI-solutions from external providers or from small start-ups which develop AI solutions. Therefore, the qualitative research results have strong links with the concept of the path dependency of companies like the Deutsche Telekom that still presents high union organizing power and a strong works council with a multi-level system. For external providers of AI-solution the results concerning workers' participation on AI-introduction may look quite different, where research has lot of to undertake in the near future.

Summary and outlook

The contextualization of the results of the qualitative case study on Deutsche Telekom underlines the importance of participation and the power resources of the respective actors as well as the role of negotiations and conflicts in the labor process and the relevance of the production and business models these are embedded in. These are key factors that help to explain AI implementation both in terms of the development of single company cases and in terms of the differences between cases. Given this, the analyzed Telekom case underlines the importance of the concept of production models. Large companies with a unionized workforce and an established multi-level system of works councils are able to offer favorable conditions for institutionalized workers' participation. In the case of the Deutsche Telekom, this condition overlapped with a tradition of social partnership that characterized labor relations and therefore conflicts in the labor process in a former public company. Based on these social relationship, management and the group works council developed new agile forms of work organization and participation to strengthen high-tech market strategies in a tough competitive environment.

At the same time, the case study underlines the importance of participation by works councils in the context of the introduction digital technologies and AI. At the Deutsche Telekom, the group works council has succeeded to develop and agree new and agile forms of participation with management as an innovative answer to AI challenges, based both on institutional power resources and the relations of social partnership with the management. This agile approach could also include a transformation process of the works council itself and a need for specific and agile-compatible qualifications of its members (Niewerth and Massolle, 2022).

Finally, in line with the concept of labor process, the Telekom case and the examples of Siemens and IBM Germany show that in ongoing technological and organizational transformation processes permanent negotiations between management and employee representatives are needed to implement agreements that adapt to deeply changing situations in employment issues. The qualitative empirical analysis has shown that corporate agreements like the "AI Manifesto" and the "digital roadmap" are able to open a road to a consensus between management and works councils to find a common way to deal with the digital transformation process of AI implementation. On the one hand, these negotiations go along also with a professionalization process of the works councils to cope with technological and organizational issues on the central level of the GWC. This centralization of qualification, competencies and capabilities to act might, on the other hand, produce a challenge within the multi-level system of employee participation to communicate such compromises of workplace democracy (Dukes and Streeck, 2023) from the central company level to the nearly one thousand works councils members on local level within the company and to advertise the political legitimacy of these labor compromises. But finally more in-depth empirical analyses are needed on the critical functioning of these company agreements on AI in the further course of time.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

TH: Writing – original draft. MW: Writing – original draft. J-TD: Conceptualization, Supervision, Writing – original draft. SS: Methodology, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The support of the Open Access Publication Funds of the Ruhr-Universität Bochum is gratefully acknowledged.

Acknowledgments

We would like to thank members of the Deutsche Telekom Group Works Council for making available several documents used for this article and also for the numerous group discussions on the topic of AI and co-determination. Without these, the preparation of this article would never have been possible.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Bahnmüller, R., Kutlu, Y., Mugler, W., Salm, R., Seibold, B., Kirner, E., and Klatt, S. (2023). *Mitsprache bei der Digitalisierung? Beteiligung von Betriebsrat und Beschäftigten in digitalisierungsaktiven Betrieben*. Düsseldorf: Study der Hans-Böckler-Stiftung.

Bargmann, H. (2022). "Gute IT digital mitbestimmen, Portrait über beschleunigte Mitbestimmungsverfahren und digitale Zusammenarbeit bei der Deutschen Telekom AG," in *Betriebs- und Dienstvereinbarungen* (Düsseldorf: Hans-Böckler-Stiftung).

Bella, N., Gamradt, J., Staples, R., Widuckel, W., Wilga, M., and Whittall, M. (2022). *Partizipation und Ungleichzeitigkeit. Eine Herausforderung für die Mitbestimmung*. Wiesbaden: Springer VS. doi: 10.1007/978-3-658-38284-1

Boyer, R., and Freyssenet, M. (2003). *Produktionsmodelle. Eine Typologie am Beispiel der Automobilindustrie*. Berlin: Edition Sigma.

Briken, K., Chillas, S., Krzywdzinski, M., and Marks, A. (2017). "Labour process and the New Digital workplace," in Briken, K., et al. (eds.). *The New Digital Workplace. How New Technologies Revolutionise Work*, eds. K. Briken, S. Chillas, M. Krzywdzinski, A. Marks (London: Palgrave Macmillan), 1–17. doi: 10.1057/978-1-137-61014-0_1

Burawoy, M. (1985). *The Politics of Production*. London: Verso.

Deutsche Telekom (2023). *HR Factbook 2022: Menschen*. Entwicklungen, Bonn: Fakten.

Doellgast, V. (2022). Strengthening social regulation in the digital economy: comparative findings from the ICT industry. *J. Soc. Econ. Relat. Work* 33, 1–17. doi: 10.1080/10301763.2022.2111987

Doellgast, V. (2023). *Exit, Voice, and Solidarity. Contesting Precarity in the US and European Telecommunications Industries*. Oxford: Oxford University Press. doi: 10.1093/oso/9780197659779.001.0001

Doellgast, V., and Kämpf, T. (in press). *Co-Determination Meets the Digital Economy: Works Councils in the German ICT Services Industry*. Paris: Entreprises et histoire.

Doellgast, V., Wagner, I., and O'Brady, S. (2022). Negotiating limits on algorithmic management in digitalised services: cases from Germany and Norway. *Transfer* 29, 1–16. doi: 10.1177/10242589221143044

Dukes, R., and Streeck, W. (2023). *Democracy at Work. Contract, Status and Post-Industrial Justice*. Cambridge: Polity press.

Edwards, R. C. (1979). *Contested Terrain: The Transformation of the Workplace in the 20th Century*. London: Heinemann.

Falkenberg, J. (2021). *Taylors Agenten. Eine arbeitssoziologische Analyse mobiler Assistenzsysteme in der Logistik*. Baden-Baden: Nomos, edition sigma. doi: 10.5771/9783748920748

Friedmann, A. (1977). *Industry and Labour: Class Struggle at Work and Monopoly Capitalism*. London: Palgrave. doi: 10.1007/978-1-349-15845-4

Gerst, D. (2020). Geschäftsmodelle mitentwickeln – ein neues Handlungsfeld der Betriebsräte. *WSI-Mitteilungen* 73, 295–299. doi: 10.5771/0342-300X-2020-4-295

Grasy, J., and Seibold, B. (2023). "Die Komplexität von KI-Systemen durch Steckbriefe bewältigen: Portrait über den Einsatz sogenannter AI-Cards bei Siemens," in *Betriebs- und Dienstvereinbarungen* (I.M.U. Institut der HBS).

Haipeter, T. (2019a). "Workers' participation: concepts and evidence," in *The Palgrave Handbook of workers' participation at plant level*, eds. S. Berger, L. Pries, and M. Wannöffel (Houndmills: Palgrave Macmillan), 149–166. doi: 10.1057/978-1-137-48192-4_8

Haipeter, T. (2019b). *Interessenvertretung in der Industrie 4.0. Das gewerkschaftliche Projekt Arbeit 2020*. Baden-Baden: Nomos, edition sigma. doi: 10.5771/9783845295770

Haipeter, T., and Schilling, G. (2023). Interessenvertretung in der Digitalisierung. Zur Bedeutung betriebspolitischer Aktivierungsprojekte der Gewerkschaften am Beispiel von "Arbeit 2020 in NRW". *WSI-Mitteilungen* 76, 392–400. doi: 10.5771/0342-300X-2023-5-392

Hirsch-Kreinsen, H. (2023a). Digitalisierung von Arbeit – ein Alltagsthema? *WSI-Mitteilungen* 76, 330. doi: 10.5771/0342-300X-2023-5-330

Hirsch-Kreinsen, H. (2023b). *Das Versprechen der Künstlichen Intelligenz*. Frankfurt, New York, Campus-Verlag: Gesellschaftliche Dynamik einer Schlüsseltechnologie.

Höfers, P., and Schröder, L. (2022). *Praxishandbuch Künstliche Intelligenz*. Frankfurt: Bund-Verlag.

Huchler, N. (2023). Arbeiten mit künstlicher Intelligenz. Wie KI Arbeit strukturiert und was sie mit der indirekten Steuerung verbindet. *WSI-Mitteilungen* 76, 365–373. doi: 10.5771/0342-300X-2023-5-365

Kötter, J., Schaffarczik, S., Daus, J. T., Repp, R., Niewerth, C., and Wannöffel, M. (2023). *Interessenvertretung unter Remote-Bedingungen: Herausforderungen und Lösungsansätze*. Mitbestimmungspraxis Nr. 55, Düsseldorf.

Krzywdzinski, M., Pfeiffer, S., Evers, M., and Gerber, C. (2022). *Measuring Work and Workers. Wearables and digital assistance systems in manufacturing and logistics*. WZB Discussion Paper SP III 2022-301. Berlin.

Kuhlmann, M. (2023). Digitalisierung und Arbeit. Eine Zwischenbilanz als Einleitung. *WSI-Mitteilungen* 76, 331–336. doi: 10.5771/0342-300X-2023-5-331

Kuhlmann, M., and Rüb, S. (2020). *Wirkungsmächtige Diskurse – betriebliche Auseinandersetzung um Digitalisierung. AIS-Studien* 13, 22–39.

Kuhlmann, M., and Voskamp, U. (2019). *Digitalisierung und Arbeit im niedersächsischen Maschinenbau*. Göttingen: SOFI Arbeitspapier, 15.

Lévesque, C., and Murray, G. (2010). Understanding union power: resources and capabilities for renewing union capacity. *Transfer* 16, 333–350. doi: 10.1177/1024258910373867

Liebold, R., and Trinczek, R. (2009). "Experteninterview," in *Handbuch Methoden der Organisationsforschung*, eds. S. Kühl, P. Strodtholz, A. Taffertshoffer (Wiesbaden: Verlag für Sozialwissenschaften), 32–56. doi: 10.1007/978-3-531-91570-8_3

Marshall, T. H. (1950). *Citizenship and Social Class: and Other Essays*. Cambridge: University Press.

Niewerth, C., and Massolle, J. (2022). *Betriebsräte in der doppelten Transformation. Ein Transferforschungsprojekt zur Organisationsentwicklung von Betriebsratsgremien*. Düsseldorf: Hans-Böckler-Stiftung.

Pfeiffer, S. (2023). "KI als Kollegin (KIK) – Repräsentative Beschäftigtenbefragung zu Künstlicher Intelligenz am Arbeitsplatz," in *Künstliche Intelligenz, Mensch und Gesellschaft, Hrsg. Norbert Huchler und Michael Heinlein* (Wiesbaden: Springer).

Rego, K. (2022). Works councils and digitalisation of manufacturing: opportunity or threat for their power position? *Econ. Ind. Democ.* 43, 1911–1933. doi: 10.1177/0143831X211054177

Rego, K., Houben, D., Brüning, S., Schaupp, S., and Meyer, U. (2021). *Mitbestimmungspraxis in der "Industrie 4.0." Möglichkeiten der Einflussnahme und Gestaltung für Betriebsräte*. HBS-Working Paper 232. Düsseldorf.

Remers, F. (2023). *Chancen nutzen, Risiken abwägen. Regelung der Nutzung von KI bei IBM von 2020 (Presentation Slides)*. WinA Transfertag 2023. Available online at: https://wina-projekt.de/veranstaltungen/a41bb0dd-aa8e-45be-8575-5067da40874a (accessed July 4, 2023).

Rüb, S., Carls, K., Kuhlmann, M., Vogel, B., and Winter, S. (2021). *Digitalisierungskonflikte. Eine empirische Studie zu interessenpolitischen Auseinandersetzungen und Aushandlungen betrieblicher Rationalisierungsprozesse*. Düsseldorf: Hans-Böckler-Stiftung.

Schäfer, M., Wannöffel, M., and Virgillito, A. (2022). Transferforschung – ein methodisches Konzept für die Analyse der Industriellen Beziehungen. *Industr. Beziehu.* 29, 129–147. doi: 10.3224/indbez.v29i2.04

Schmalz, S., and Dörre, K. (2014). Der Machtressourcenansatz: Ein Instrument zur Analyse gewerkschaftlichen Handlungsvermögens. *Industr. Beziehu.* 21, 217–237. Available online at: http://www.jstor.org/stable/24330817

Thompson, P. (1990). "Crawling from the wreckage: the labour process and the politics of production," in *Labour Process Theory*, eds. D. Knights, H. Willmot (London: Palgrave Macmillan), 95–124. doi: 10.1007/978-1-349-20466-3_3

Thompson, P., and Laaser, K. (2021). Beyond technological determinism: revitalising labour process analyses of technology, capital and labour. *Work Global Econ.* 1, 139–159. doi: 10.1332/273241721X16276384832119

Ver.di, IBM, BMAS. (2020). *Künstliche Intelligenz: Ein sozialpartnerschaftliches Forschungsprojekt untersucht die neue Arbeitswelt*. Available online at: https://www.verdi.de/++file++5fc901bc4ea3118def3edd33/download,\20201203_KI-Forschungsprojekt-verdi-IBM-final.pdf (accessed July 4, 2023).

Wilkinson, A. (2011). *The Oxford Handbook of Participation in Organizations*. Oxford: Oxford University Press. doi: 10.1093/oxfordhb/9780199207268.001.0001

Frontiers in
Artificial Intelligence

Explores the disruptive technological revolution of AI

A nexus for research in core and applied AI areas, this journal focuses on the enormous expansion of AI into aspects of modern life such as finance, law, medicine, agriculture, and human learning.

Discover the latest Research Topics

See more →

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact