Comes *EPJ Data Science* (2026) 15:42
https://doi.org/10.1140/epjds/s13688-026-00634-5

EPJ.org

EPJ Data Science

RESEARCH

Open Access

# Sensemaking AI: Introducing a research and design agenda for human–AI networks

Tina Comes[1,2*]

Handling Editor: Sachit Mahajan

*Correspondence:
t.comes@tudelft.nl
[1]Faculty Technology, Policy & Management, TU Delft, Delft, The Netherlands
[2]Institute for the Protection of Terrestrial Infrastructures, DLR, Cologne, Germany

**Abstract**

Digital technologies and AI promise to optimise complex systems through data-driven decisions, predictive modelling, and anticipatory action. However, this optimisation imperative creates a fundamental paradox: as systems excel at achieving measurable objectives, they may erode the collective intelligence and adaptive capacity of our societies. Recognising this tension, the field of Human-Centred AI (HCAI) has emerged to develop design principles such as explainability, fairness, and transparency to ensure that AI aligns with human values. However, research on HCAI often focuses on idealised interactions, neglecting the pressure, moral dilemmas, and social dynamics typical of today's complex problems. This paper introduces and advocates for a paradigm shift towards *Sensemaking AI*: AI that supports collective meaning-making processes in evolving human-AI networks. This novel perspective recognises that algorithmic and AI systems actively participate in the social processes through which humans interpret information, coordinate responses, and adapt their values. Grounded in sensemaking and decision theory and informed by a scoping review of the HCAI literature, this paper identifies three connected research areas: (i) sensemaking-aware automation that preserves interpretive flexibility; (ii) collective agency for network-level control; and (iii) value-aware sensemaking that supports collective meaning-making. These principles form the basis for Sensemaking AI as a design and research agenda that prioritises collective meaning-making and democratic deliberation in networks.

**Keywords:** Scoping review; Sensemaking AI; Human-AI interaction; Decision theory; Human-Centred AI; Complex systems; Collective intelligence; Optimisation: Human-AI networks

## 1 Introduction

Our societies are complex dynamical systems increasingly shaped by digital technologies. As decision-makers grapple with increasing complexity, AI is portrayed as a solution for its ability to process large amounts of data and reduce the cognitive biases in human decisions by providing data-driven answers. For instance, the UN Office for Disaster Risk Reduction (UN DRR) opens its Tech4DRR report with "*AI technologies enable the analysis of vast amounts of data, uncovering patterns and insights beyond human capacity*" (UNDRR [122]). Digital twins (Fan et al. [35]), automated decisions (Coppi et al. [22]), and predic-



Springer

tive models for 'anticipatory action' (Kjærum & Madsen [65]) promise efficient, scalable, rapid and cheap solutions to address complex and time-compressed problems.

With the promise of 'taming complexity' (Gil et al. [42]), AI has become both a tool to understand complex systems and a driver that shapes their behaviour. As such, AI systems do not just optimise a system: they shape the informational and social environments via which humans understand their environment and respond to the challenges they are confronted with. Recognising this dual nature, this paper echoes the calls for *Machine Behaviour* (Rahwan et al. [96]), *Human-Centred AI* (HCAI) (Ozmen Garibay et al. [90]) and *Hybrid Intelligence* (Akata et al. [3]) that all conceptualise AI as a social-technical system. Especially HCAI has emerged as a prominent field advocating for explainable, fair, and transparent AI systems that keep humans "in the loop" rather than replacing them (Shneiderman [107]). This human-centred approach seeks to address the optimisation paradox by designing AI that enhances rather than diminishes human agency. By explicitly putting forward dedicated design principles, such as explainability, fairness, trust, transparency, HCAI provides a lens to examine whether current research addresses the social, collective dimensions of meaning-making in complex systems.

Yet HCAI is based on idealised situations that insufficiently address the time pressure, high stakes, volatility, moral dilemmas and shifting networks typical for decisions in today's complex problems – even though these conditions have been shown to fundamentally change information processing and lead decision-makers to overlook, discard, or not act upon information (Paulus et al. [93]). In complex problems, where multiple interconnected challenges demand rapid coordination across scales and domains (Mahajan et al. [77]) this lack of understanding may lead to overlooking important implications of the use of AI, which in turn jeopardises the central tenet of meaningful human control (Cavalcante Siebert et al. [12]).

*Sensemaking*, defined as process by which humans collectively interpret ambiguous situations, construct meaning from uncertain information, and adapt as understanding evolves (Weick [133]), offers a crucial lens and theoretical framework to address this tension. Unlike optimisation approaches with predetermined objectives, sensemaking embraces the continuous social processes through which values and objectives emerge and evolve. This is essential when dealing with complex problems that can only be addressed by decentralised interaction rather than top-down optimisation.

This paper contributes in three ways to the ongoing discourse on optimisation and AI: building on foundations in sensemaking support systems (Muhren & Van de Walle [83]; Seidel et al. [102]) and decision theory (French [39]), it introduces *Sensemaking AI* as *AI that actively supports sensemaking processes in evolving human-AI-networks*, and delineates it from other paradigms such as HCAI (Shneiderman [107]) or machine behaviour (Rahwan et al. [96]). Unlike AI for collective intelligence (empirically measuring or optimising group performance for dedicated tasks in stable teams) or Human-Cyber-Physical Systems (emphasising coordination requirements in *predefined* configurations), Sensemaking AI centers on sustaining the interpretive flexibility and democratic deliberation through which values and objectives *emerge* in uncertain and ambiguous situations within evolving networks, where both tasks and team compositions are not known a priori.

Empirically, a scoping review of 101 academic articles from the Human-Centred AI (HCAI) literature examines which principles dominate current research and how these principles address (or overlook) the dynamics of human-AI interaction in complex, time-

sensitive, and morally fraught environments. Poly-crises serve here as illustrative cases to highlight the limitations of optimisation-centric AI design. Poly-crises represent an overlooked class of problems characterised by the intersection of complexity; uncertainty and ambiguity; and time pressure. These characteristics are increasingly common, yet systematically absent from HCAI research, and I argue that poly-crises reveal fundamental limitations in optimisation-centric AI design.

Programmatically, the paper then develops design principles and a research agenda on Sensemaking AI along three axes: sensemaking-aware automation that promotes interpretive flexibility; collective agency for network-level control; and value-aware Sensemaking AI. As such, this paper is meant to inspire the reader and start a conversation across computer, cognitive and behavioural sciences that acknowledges that fundamentally we humans continuously engage in the social process of making sense of our environment (Maitlis [79]).

## 2 On the impact of AI on Sensemaking & decision-making

This section provides the theoretical foundations for understanding how AI transforms human sensemaking and decision-making. I first provide insights into Sensemaking as a theory and process for navigating complexity. Then, I explore the foundations of Human-Centred AI to showcase how the notion of control is shifting with the ubiquity of AI and the complexity of the problem addressed. From there, I move to sketch how AI systems reshape information-decision feedback dynamics in networked social-technical systems.

### 2.1 Sensemaking

We are confronted with several accelerating poly-crises (Søgaard Jørgensen et al. [110]), i.e., situations, where multiple, interconnected crises, including climate change, geopolitical instability, poverty and economic inequality, interact and amplify each other. This class of problems is inherently "wicked", i.e., poly-crises are characterised by contested problem definitions and ambiguity, interdependencies, and evolving values (Rittel & Webber [99]). A further compounding factor is the urgency of poly-crises, where the perception of 'time running out' compresses deliberation time. Data-driven optimisation and AI promise clearly identifying the 'best' alternatives even– or especially – if problems are complex and time is short. However, by definition, wicked problems cannot be solved by traditional optimisation approaches (French [40]). I argue that the question is not how to optimise predefined objectives, but how to collectively construct what problems mean, the core of Sensemaking Theory.

Weick's theory of sensemaking (Weick [132, 133]) provides the foundations to explain how humans navigate such ambiguity and uncertainty. Sensemaking is the process of meaning making, by which humans structure and process the stream of unfiltered, chaotic data, and turn it into meaningful and actionable information (Weick [132]). Sensemaking is a creative process, wherein we humans construct 'bridges' to address uncertainties consisting of ideas, thoughts, emotions, feelings, and memories (Sharoda & Reddy [105]). Importantly, sensemaking is a social process, through which decision-makers interact with their peers and the environment, they coordinate, and receive feedback through enactment, allowing the formation of collective action agendas (Maitlis & Christianson [80]). Sensemaking also entails the process of identity construction, by which humans come to understand what is 'meaningful' in their own identities, teams and organisations (Helms Mills et al. [49]).

Unlike rational decision-making models that assume clear problems and predetermined objectives (Gralla et al. [44]), sensemaking recognises that in complex environments, actors first construct what situations mean before determining their responses (Comes, Van de Walle, et al. [18]). Where optimisation assumes fixed objectives and measurable trade-offs, sensemaking analyses how those objectives emerge through interaction, feedback, and evolving understanding. Rather than converging on stable solutions, sensemaking sustains the ambiguity necessary for creative reinterpretation and collective learning (Weick & Sutcliffe [131]).

As AI systems increasingly filter information, prioritise, and suggest what to do, they inevitably become participants in these sensemaking dynamics and shape the cognitive and social foundations on which sensemaking depends. The question that emerges is not simply how to design AI that supports human decisions, but how to understand AI's role in shaping the very processes through which our collective understanding and coordinated action emerge.

### 2.2 Human-centred AI: towards control in human-AI networks

The term human-centred AI is increasingly popular in response to the many concerns about AI risks and human agency in increasingly digitally mediated environments (Capel & Brereton [11]). Instead of a world in which AI optimises our lives and takes control, human-centred AI presents a design paradigm to ensure that AI serves people and enhances human capabilities rather than replacing them (van Berkel et al. [126]). By definition, such an AI is trustworthy, reliable and safe for humans to use (Auernhammer [6]; Shneiderman [107]).

Adjacent research streams examine human-AI interaction, including interaction of larger groups, from different angles, especially AI for Collective Intelligence (Riedl & De Cremer [97]) and Human-Cyber-Physical Systems (Lou et al. [75]). Yet, the explicit focus of Human-Centred AI on normative design principles makes it the appropriate lens for this scoping review. Collective intelligence is defined as the shared problem-solving ability that arises from the interaction and combined efforts of a group through collective memory, attention, and reasoning (Woolley et al. [136]). Obviously, AI for collective intelligence focuses on how *AI* can augment the capacity of human groups through information collection, coordination, and decision support for distributed collectives (Gupta & Woolley [45]). Human-Cyber-Physical Systems research, emerging from manufacturing contexts, examines the operational integration of human cognition, cyber systems, and physical assets (Lou et al. [75]) and emphasies architectures, verification methods, and system-level coordination, focusing on human-in-the-loop, human-on-the-loop, and human-in-society paradigms as engineering requirements. Both areas are organised around capabilities and group performance and coordination, rather than *normative* principles about how AI should relate to human values, control or agency. In contrast, HCAI is explicitly organised around normative design principles that articulate how AI should support human interests (Shneiderman [107]), providing an ideal starting point for examining how tension between optimisation and emergence.

Central to HCAI's normative commitments and design principles is the notion of meaningful human control. However, what "control" means shifts fundamentally as AI becomes embedded in complex socio-technical systems. Often, control is framed as a question of who carries out which task or holds responsibility for it. The optimal distribution of tasks

between humans and machines has been a central question since the 1950s, when Fitts [36] introduced a protocol to decide which tasks are better performed by machines or by humans in the context of air traffic control. Today, this question has evolved to address how AI can support and optimise human decisions in complex problems while avoiding the pitfalls of cognitive or motivational biases. Tasks are assigned to humans or machines based on their respective expertise to optimise performance (Tausch & Kluge [118]), with humans being better at tasks requiring creative and social intelligence (Ponti & Seredko [95]). If tasks are allocated to machines, rapidly questions become acute about human autonomy and control (Abbass [1]).

However, the question of control in human-AI systems extends beyond task allocation. The question is not merely about *who* (or what) decides. Rather, we need to ask how we can preserve meaningful human agency when AI systems shape the interpretive context in and through which we make choices. As AI filters, structures, and sequences information, it influences attention, salience, and relevance, thereby directing what we read, see, ignore, or find important. In doing so, AI systems (co-)construct the conditions in which human sensemaking unfolds.

Many of the frameworks to understand control stem from understanding human operators of machines, ranging from submarines (Sheridan et al. [106]) to air traffic (Council [23]). These frameworks put forward assume a single human operator interacting with one machine (Endsley [32]; Parasuraman et al. [91]). Yet, addressing complex challenges requires decentralised approaches, where many people (and potentially algorithms) work together (Mahajan et al. [77]). Increasingly, there is a discussion around human-AI teams (HAT), "*a purposeful combination of human and cyber-physical elements that collaboratively pursue goals that are unachievable by either individually*" (Alix et al. [5]). The literature distinguishes configurations by the number and interaction patterns of humans and AI systems. Human-AI teams typically are composed of multiple, but few people and machines (Endsley [33]). Other control configurations include a single AI system that supports many humans (e.g., one AI-based early warning system that serves large populations) or autonomous AI systems that operate without or with minimal human oversight (e.g., an autonomous robot for urban search and rescue). Complex problems, however, require the coordination of large-scale, ad-hoc and decentralised networks of humans and AI. Figure 1 summarises the evolution of research on human-machine interaction and control. With increasingly complex challenges and pervasiveness of AI, research needs to shift from AI that mimics human reasoning and single-operator systems to decentralised evolving human-AI networks that require coordination and human control.

While many AI-supported decisions are intended as quick fixes and short-term optimisations, they often produce lasting consequences. Once enacted, decisions alter social networks, infrastructures, and expectations, creating path dependencies that narrow future choices (Webster [130]). These effects are especially problematic when they shape collective sensemaking trajectories (Helms Mills et al. [49]), influencing how future situations are interpreted and what actions seem legitimate. Moreover, AI systems themselves evolve through feedback: large language models, for instance, adapt to user behaviour, reinforcing patterns of engagement and attachment (Kim et al. [64]) and producing recursive information bubbles (Jacob et al. [57]). Over time, both the humans and the AI intended to support the humans may drift away from the original intentions in ways that become difficult to detect, control or reverse. Preserving human control thus requires more than

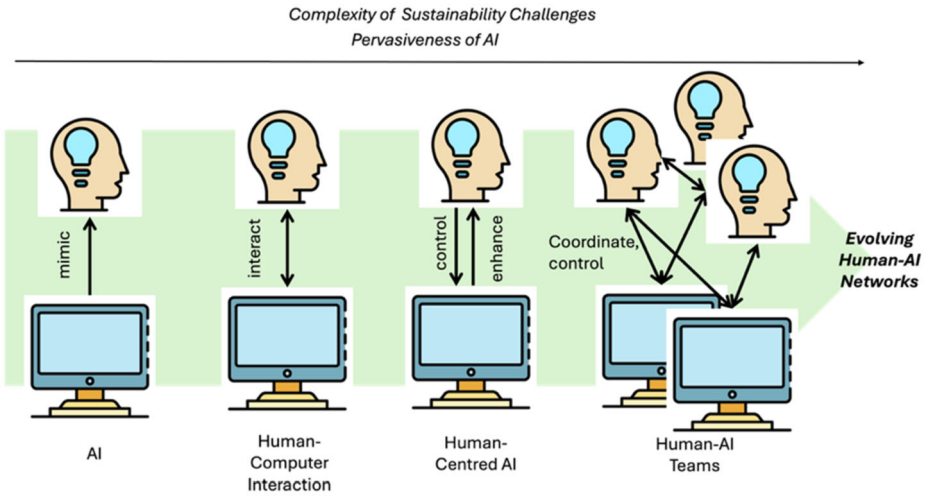


**Figure 1** The need to understand evolving human-AI networks

oversight: it demands mechanisms for *temporal reflexivity*: the ability to recognise deviations from purpose and intervene in unfolding trajectories when needed.

These temporal shifts and the decentralised nature of human-AI networks challenge conventional notions of agency and control, particularly as AI becomes embedded in systems that evolve faster than human governance can follow. In the next section, I examine how these dynamics manifest via information-decision-feedback loops.

### 2.3 Optimisation cascades and information-decision-feedback

AI and humans exist in a complex control relationship: AI influences human sensemaking and decision-making, and humans engineer, train, steer, and – possibly – control the AI (Rahwan et al. [96]). Despite this interdependence, information and decisions are typically studied independently, and decision-information-feedback is only marginally considered.

By their very nature, poly-crises require urgent interventions despite their complexity. This urgency matters, as time compression alters information processing, information sharing, and human decision-making behaviour. With prospect theory in the 1970s (Kahneman & Tversky [61]), the need to include cognitive aspects in risky decisions became prominently recognised and inspired a wealth of research pointing to the cognitive and motivational biases that they bring (Klein et al. [66]; Paulus et al. [93]; Weick [132]). An obvious question then is: can AI improve human decision-making by overcoming the cognitive or motivational biases?

To unpack this question, an analysis of human decision-making processes is helpful, see Fig. 2, left side. The human process is driven by the interplay of sensemaking, coordination and decision-making, by which humans adapt to the changing informational and social environment. While often, it is assumed that decision-making is the guiding principle to organise coordination, research has shown that it is the interaction of sensemaking and decision-making (Comes, Van de Walle, et al. [18]; Gralla et al. [44]). This implies that as the understanding of the situation changes, also the objectives, preferences, values and thereby the required or desired solutions change, fundamentally contradicting the linear problem solving paradigm, by which first a problem is formalised, and then solved (Volkema [129]).

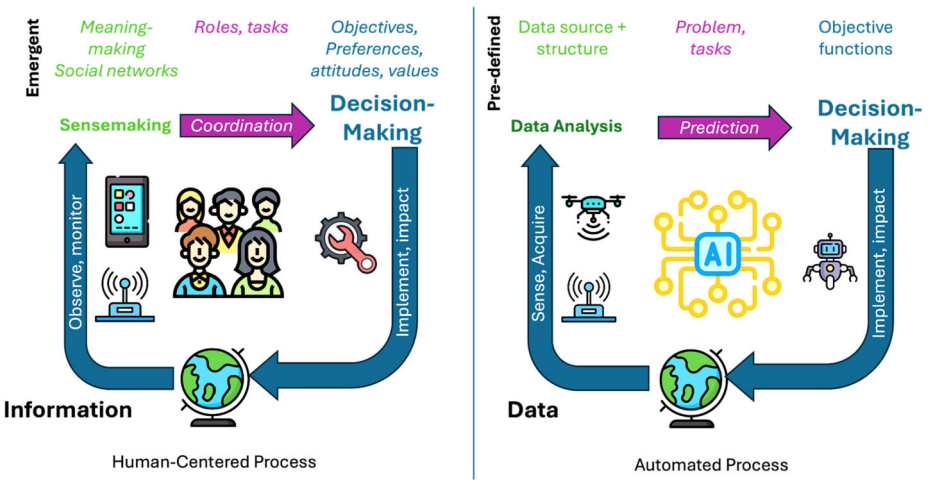


**Figure 2** The cycle of sensemaking and decision-making. Comparing human-centred (left) and automated processes (right)

AI and optimisation algorithms fundamentally change how we perceive our environment, with whom and how we interact to find solutions (Kiesler & Sproull [63]). For illustrative purposes, Fig. 2 (right side) shows the most extreme case where one or several AI systems autonomously acquire and analyse data; optimise, decide and implement the decision in '*optimisation cascades*'. This is not meant as a prescription, but to contrast with the human sensemaking cycles. In this case, the creative, social and identity-forming process of sensemaking is replaced by data analysis. Optimisation cascades manifest across scales. At the individual level, recommendation algorithms and LLMs optimise for engagement based on personal information, shaping what people read or see and thereby gradually shifting user preferences and choices (Sharma et al. [104]). At the policy level, performance metrics drive decisions under the umbrella of New Public Management, even if they only poorly capture intended outcomes. This phenomenon that Muller [84] called the "tyranny of metrics" *distorts* our understanding of complex phenomena such as poverty, climate change, or welfare.

A fundamental tension between human sensemaking cycles and optimisation cascades is that human processes are iterative social cycles where problem definitions and objectives co-evolve with shared understanding. In contrast, optimisation processes replace the underlying creative ambiguity with predetermined static metrics and automated execution of the optimal decision. The risk here is not only that machines make "wrong" decisions, but that the optimisation fundamentally impacts human sensemaking, shaping social orders, norms of our interactions, and what we perceive as 'good' or even thinkable solutions.

Today, hybrid approaches, by which many humans and machines work together, are becoming the norm. An open question is how to coordinate the resulting dynamic networks of humans and AI, while controlling information-decision feedback loops. Here, coordination is defined as the set of procedures by which teams plan, organise, orchestrate and integrate their activities to achieve shared goals (Malone et al. [81]). As such, coordination entails activities such as information sharing, planning, task allocation or scheduling (Neale et al. [87]). As AI is becoming increasingly prevalent, coordination entails human-

human coordination (e.g., information sharing in teams), human-AI coordination (e.g., task allocation between humans and AI systems), and increasingly, AI-AI coordination (e.g., multi-agent systems). Especially for agentic AI systems, the challenge shifts from coordinating humans who use AI tools toward coordinating hybrid networks where both humans and AI systems act as semi-autonomous agents.

For human actors, several studies have confirmed that decision performance on distributed tasks improves if individuals know who has access to what information (Stasser & Titus [113]; Stewart & Stasser [114]), which is problematic in complex networks and with black box AI models. Further, if AI – and especially generative AI – optimises information based on past interactions, it can introduce or reinforce existing biases (Sharma et al. [104]), which are then amplified via path-dependencies in sensemaking and decision trajectories. Given the broad impact of AI and optimisation on human sensemaking and decision-making in complex networks, the next sections examine how and in how far current Human-Centred AI literature addresses these fundamental challenges.

## 3 Methods

To assess how the literature navigates the tension between data-driven optimisation and human meaning-making in uncertain and evolving contexts, I focus on the HCAI literature. HCAI addresses the relationship between human agency and AI systems. As such, it provides an important lens for evaluating how AI design accounts for the cognitive, social, and ethical dynamics of complex decisions. To explore this evolving body of work, I conducted a scoping review. Scoping reviews are designed to determine the scope or coverage of an emergent body of literature (Munn et al. [85]) and identify gaps in emerging interdisciplinary fields such (Sadek et al. [100]). The scoping review was conducted on rayyan (https://www.rayyan.ai/) and followed the extended PRISMA protocol for scoping reviews (Tricco et al. [121]). Figure 3 summarises the screening and analysis process.

### 3.1 Article identification

The search was based on Scopus, because of its comprehensive coverage of interdisciplinary sources. To determine how human values and decisions are represented in the HCAI literature, the search focused on human-centred AI, searching combinations of "human-centred AI", "human-centered AI" OR "HCAI" with "Decision support", "Values" OR "Principles". This framing focuses on decision problems to reveal the limitations of optimisation-driven AI. The search was limited to articles in English, focusing on original manuscripts, case studies and perspective papers. To account for the rapid growth of the field, a preliminary pilot was conducted in June 2024, and the final search was executed on February 14, 2025. Excluding duplicates, the search led to 419 articles.

### 3.2 Screening

The titles, abstracts and keywords of all articles were screened. Articles had to focus on supporting specific decision-making or sensemaking tasks and discuss the distribution of the tasks among humans and machines under the framework of human-centred AI to ensure that the different HCAI concepts or principles could be mapped out. This initial screening process resulted in 185 papers that were retained for the full-text review.

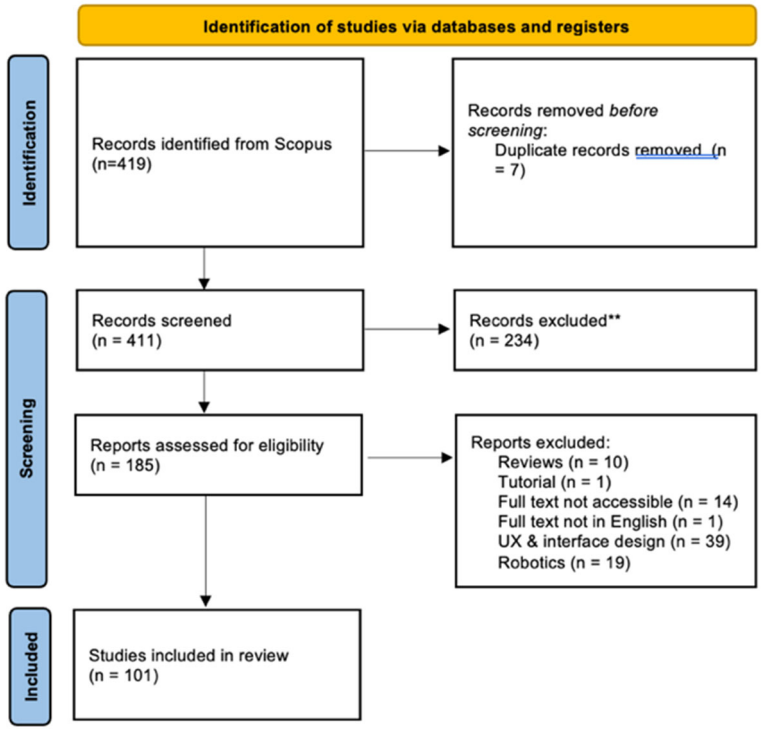


**Figure 3** PRISMA-Scoping Review Protocol followed for this study

### 3.3 Full-text eligibility

The articles were then analysed based on the extent of discussion on human-AI interaction, decision-making/ sensemaking and HCAI concepts. Studies were excluded if they: (i) focused on physical interaction (e.g., robotics); (ii) addressed only interface design without engaging HCAI principles; (iii) referred only to conventional statistics, not AI; or (iv) were categorised as reviews or tutorials, or (v) were inaccessible, not in English, see Fig. 3. This led to 101 papers that were retained for the analysis. An overview of all included papers is provided in Annex 1.

### 3.4 Analysis and extraction framework

The analysis began by mapping publication trends and methods across the reviewed studies. For the *methods* used, I categorise (i) perspective and conceptual papers that theorise or aim to guide the use of AI; (ii) empirical studies that observes how humans interact with machines in real life; (iii) behavioural experiments; (iv) design studies that build and implement the AI, also including technical studies, and (v) surveys. Further, I distinguish the *field*, for which the AI is designed, from the *type of AI or algorithm*. I categorise: (i) *machine learning* (*ML*), referring to traditional supervised and unsupervised learning approaches (classifiers, regression models, etc.) for classification or prediction, explicitly *excluding* generative methods; (ii) *generative AI*, encompassing large language models and other systems that generate novel content; (iii) *natural language processing* (*NLP*), when papers emphasise language understanding and processing capabilities; (iv) *agentic AI*, referring to systems characterised by goal-directed, adaptive behavior and autonomous action capabilities, most often manifested in conversational agents; and (v) *AI agnostic/unspecified*,

**Table 1** Overview of the Analysis & Extraction Framework for the Scoping Review

| Dimension | Categories |
|---|---|
| Research Method | Perspective/conceptual; qualitative empirical; behavioural experiments; design and tool developments; surveys |
| Field of Application | Medicine & Health; Public policy & governance, Business; Education; Crisis & Safety; Manufacturing; Software Engineering; Transport; Other |
| Type of AI | AI agnostic/unspecified; Machine Learning, Natural Language Processing (NLP); Generative AI; Agentic AI (goal-directed adaptive system) |
| Role of Human And AI & control | Single-user support (1 AI:1 human); autonomous AI (1 AI: minimal human intervention); Human-AI teams (n AI: m humans, typically with coordination); One-AI-many-humans (1 AI: m humans) |
| HCAI DESIGN Principles | Explainability, Fairness, Trust, Transparency, Accountability, Solidarity, Contextualisation, Empowerment, Safety, Humanity, Control, Agency, Privacy, Useability, Responsibility, Automation, Equity, Situational Awareness, Bias, Sustainability |

when papers do not commit to or discuss specific algorithmic approaches. Papers were coded based on the principal algorithmic approach that the paper used for positioning. In two cases where papers explicitly engaged with multiple AI types, both were recorded.

From there, I investigate the *role* of the human vis a vis the machines to understand how and in how far humans are – indeed – central to HCAI, and what the implications are for optimisation, sensemaking and decision-making. Questions are who has decision authority (the human or the machine); and how many humans are interacting with how many machines. I conceptualise this in terms of the following categories: (i) *single-user support*, where one AI supports an individual human decision-maker who is responsible for the decision; (ii) *autonomous AI*, where the AI makes the decision autonomously with minimal human oversight; (iii) *Human-AI teams*, where multiple humans collaborate with one or more AI systems in a coordinated task structure; or (iv) *one-AI-many-humans*, where a single AI system (e.g., a recommendation algorithm or chatbot) interacts with multiple humans who are not necessarily coordinated as a team.

Importantly, the *type of AI system* refers to the algorithmic approach or technological characteristics while *role* in *human-AI relations* describe the control structure, i.e., who or what holds decision-making authority and how many humans interact with how many AI systems. For instance, an agentic AI system (type) might operate under human control in a single-user support configuration, or it might be deployed with autonomous decision-making authority. These dimensions are analytically independent.

Finally, the review assessed which *HCAI design principles and concerns* are most prominent in the literature, and how they co-occur. Because many terms in the literature—such as fairness, explainability, and automation—blur the boundary between normative values and technical features, I treat them collectively under the umbrella of *HCAI design principles*. This includes both ethical principles (e.g., solidarity, equity) and system-level properties (e.g., transparency, control, privacy). The initial coding categories were established on the basis of major guidelines (EC [30]; OECD [89]; UNESCO [123]), as well as the seminal HCAI literature (Shneiderman [107]). Through bottom-up coding concepts were added. An analysis of frequency and co-occurrence patterns was then used to explore dominant principles and gaps in the HCAI literature.

Table 1 provides the coding framework indicating the dimensions of analysis and categories used. The coding was conducted by the author as a single coder. The coding framework presented in Table 1 served as the explicit coding scheme for consistent categorisa-

tion across all 101 papers. As with any single-author review, the possibility that individual papers could be assigned to alternative categories cannot be excluded. For HCAI design principles, papers were coded with all applicable concepts, resulting in multi-label assignments (mean 2.7 principles per paper; 84/101 papers engage with two or more principles), which enables the co-occurrence analysis presented in Sect. 4.3. For field, AI type, and role of human-AI and control, papers were coded along their primary characterisation for ensure analytical clarity. In a small number of cases where papers explicitly addressed multiple fields (4 papers), AI types (2 papers), or human-AI role configurations (3 papers), multiple labels were assigned. While this primary-characterisation approach may underestimate connections and nuances for individual papers, the aggregate patterns across 101 papers remain robust. The multi-label coding of HCAI design principles as the central analytical dimension of the paper captures the conceptual intersections.

### 3.5 Design principles for Sensemaking AI

In addition to the scoping review, this paper develops a set of theoretical arguments and proposes three interconnected design principles for Sensemaking AI with the aim of starting to address conceptual blind spots of the current literature. The design principles proposed are developed through an abductive process (Sætre & Van de Ven [101]) that iteratively links theoretical framing from Sensemaking and Decision-Making theories and empirical patterns from the review with illustrative examples from poly-crises to develop a research and design agenda for AI that supports meaning-making in complex systems.

## 4 Results

This section presents the results of the scoping review and examines how HCAI research engages with the tensions between optimisation, human agency, and sensemaking in complex environments. By analysing patterns in methods, applications, AI system types, roles of humans and AI, and HCAI design principles, this results section aims to provide insights into how the field has evolved and where important gaps may remain.

### 4.1 The how, what and where of HCAI

Along with the rise of AI, there is also growing interest at the intersection of human-centred AI and decisions sparked by the first HCAI publication by Shneiderman [107]. Figure 4 shows the distribution and temporal evolution of research methods and AI types across the HCAI literature, revealing both the diversity of approaches and important shifts over time. The figure shows a maturing field shifting from algorithm-independent theorising towards increasing engagement with specific AI systems. Initially, AI-agnostic papers dominated the discourse (with more than 60 % of publications in 2019-2021), as researchers established foundational HCAI principles independent of the actual implementation. These publications theorise or analyse '*AI*' as a generic term, being agnostic of the specific algorithm (Akula & Garibay [4]; Bingley et al. [9]; Hoch et al. [52]).

This pattern shifted dramatically in 2024: AI-agnostic papers dropped to about 7 % (2/28 papers), while traditional ML surged to 50% (14/28 papers), and agentic AI is emerging prominently at 25% (7/28 papers). This shift suggests HCAI is moving from abstract principles toward examining how different AI algorithms, with their distinct capabilities and limitations, impact human agency and control. Machine learning (ML), including deep learning, is thus the most common approach with more than 35 % of publications, either alone or in combination with other methods such as natural language processing

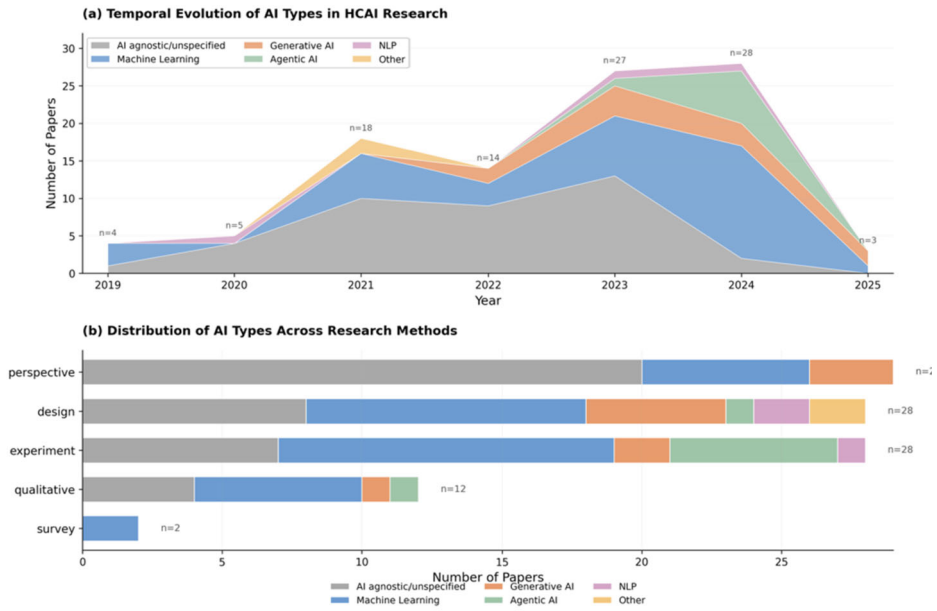


**Figure 4** Distribution and evolution of AI types and research methods in the HCAI literature. Top panel (a) shows the temporal evolution of HCAI papers and the different types of AI that are discussed, highlighting the evolution from AI-agnostic papers to more specific applications. Bottom panel (b) shows the types of AI chosen for different research methods, highlighting that perspective papers most often are algorithm-agnostic

(NLP). Increasingly, generative AI papers are a part of the HCAI literature (Buçinca [10]; Erlei [34]; Kattnig et al. [62]). Agentic AI, which refers to AI systems designed with goal-directed, adaptive behaviour and the capacity for autonomous action within bounded environments (Acharya et al. [2]), is also increasingly well-represented with almost 10% of publications, indicating a growing interest in AI autonomy and adaptive behaviour, most often via humans interacting with an autonomous agent (Criscuolo & Dolci [24]; Gou et al. [43]).

The distribution of research methods (part b of Fig. 4) shows a diverse set of research methods, in which experimental, design and perspective approaches dominate, each with almost 30 papers. Under design methods, papers are classified that design, build, and test Human-Centred AI applications, e.g., (Elahi et al. [31]; Erlei [34]; Sun [117]), indicating a focus on development and evaluation. In contrast, the many perspective and opinion papers emphasise theoretical and conceptual discussions, primarily regarding the value perspective in HCAI, especially in sensitive contexts such as education, health or crisis management (Comes [17]; Kattnig et al. [62]; van Leersum & Maathuis [127]). Among these 29 perspective papers, 20 (69%) discuss HCAI principles without specifying algorithmic approaches, suggesting these authors aimed to establish principles transcending particular technical implementations. Controlled experiments often study cognition and behaviour by focusing on how users interact with an AI. These studies link HCAI to Human-Computer-Interaction and often test for the impact of specific principles such as fairness or explainability (Flathmann et al. [37]; Gajos & Mamykina [41]). These experimental studies engage most often with traditional ML (more than 41%), likely reflecting the controlled settings needed for behavioural research, while design studies show more diversity across AI types. Qualitative methods that study the use of AI in situ and sur-

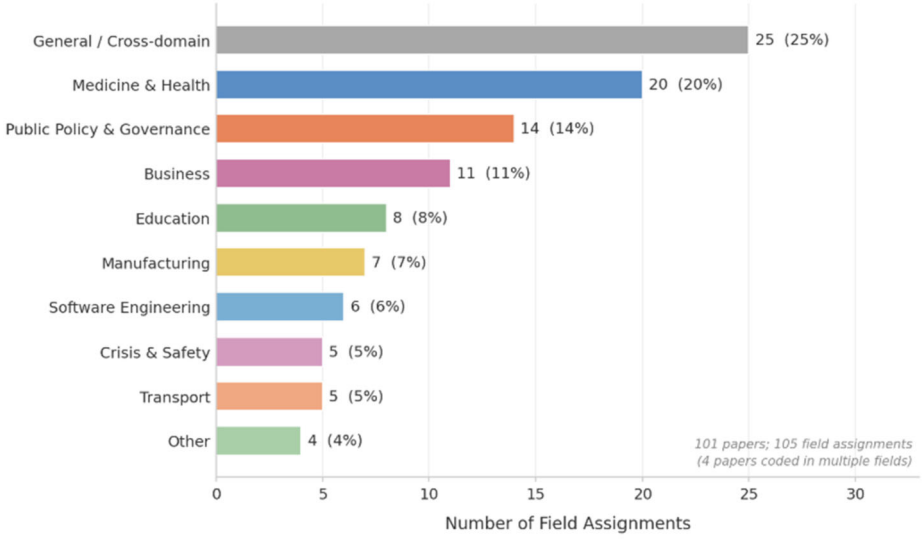


**Figure 5** Distribution of HCAI research across application domains. For the papers that discuss a domain, Medicine & Health dominate with 20 papers (20%), followed by Public Policy & Governance (14%), Business (11%) and Education (8%). Fields such as Crisis & Safety, Manufacturing, Software Engineering and Transport have received less attention. The diverse 'Other' category (4%) includes law, agriculture, and language, demonstrating HCAI's broad interdisciplinary reach

veys that focus on perception and use of technology especially in work processes are less common (Bingley et al. [9]; Herrmann & Pfeiffer [50]).

The review also shows that HCAI has made its entrance in a diversity of fields, cf. Fig. 5. Applications in medicine and health are most prominent (20/101 papers), most often in the context of clinical decision support systems for diagnosis or treatment e.g., (Lee et al. [68]; Van Berkel et al. [125]; Verma et al. [128]). This is followed by public policy and governance (Lee et al. [69]; Lettieri et al. [70]; Stapleton et al. [112]) including several papers that discuss sustainability aspects (Sigfrids et al. [109]). Business and managerial applications (Freire et al. [38]; Hoch et al. [52]) follow. In education (Chaudhry et al. [13]; Duan et al. [29]) and crisis management, papers also discuss the risks of AI especially regarding the introduction of new biases (Chaudhry et al. [13]). 'Other' is a highly diverse category, including law, agriculture, and language.

### 4.2 The role of AI in HCAI

When it comes to the roles of the AI, Fig. 6 shows that with 66/101 papers, the large majority of papers focus on an AI supporting an individual decision-maker, ranging from supporting elite sports coaches (Maiden et al. [78]) to supporting elderly app users in smart cities (Elahi et al. [31]), from nurses steering patients to a hospital (Li et al. [72]) to managers making strategic decisions (Passlack et al. [92]). Fewer papers (10/101) discuss Human-Centred AI for autonomous decision-making by which the AI fully automates decisions (Bingley et al. [9]; He et al. [47]; Jin et al. [60]; Lee et al. [69]; Nabizadeh Rafsanjani & Nabizadeh [86]; Shulner-Tal et al. [108]; Suchan et al. [116]; Yazdanpanah et al. [137]). Despite the increasing prevalence of AI in society, only nine papers discuss one AI that supports many humans, primarily in the context of crowdsourcing (Sprenkamp et al. [111]). Further thirteen papers study human-AI-teams, and the implications for group

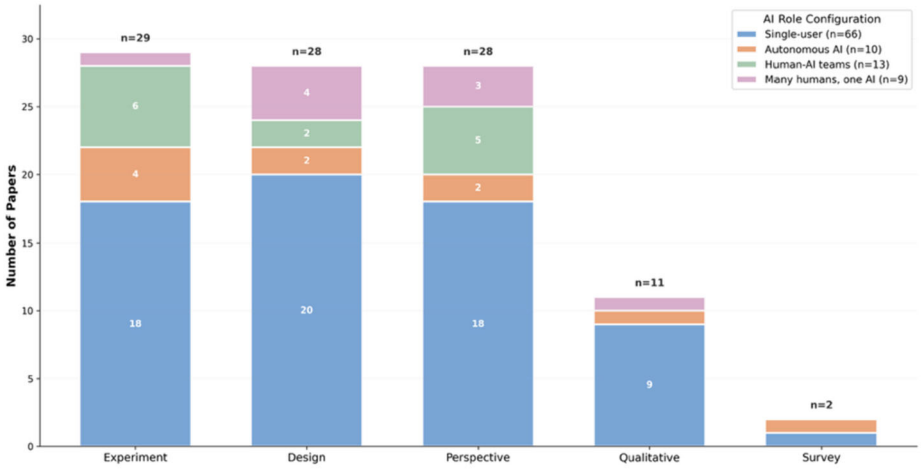


**Figure 6** Stacked bar chart shows absolute values of no. papers per research method. Experimental studies (64%), design studies (71%), perspective papers (63%), and qualitative research (80%) all predominantly examine dyadic human-AI interactions, with limited attention to networked configurations (22% overall). This pattern demonstrates that the gap in addressing collective sensemaking in human-AI networks is systematic across all methods

dynamics (Bansal et al. [7]; Cooke et al. [21]; Flathmann et al. [37]; Riedl [98]), for instance in medical teams (Hagemann et al. [46]; Verma et al. [128]) or manufacturing (Hoch et al. [52]).

Research methods may shape which aspects of human-AI interaction receive attention. Figure 6 shows the proportion of AI roles for different research methods. Experimental studies (n = 29) and design studies (n = 28) predominantly examine scenarios where AI supports a single-user (18/29, 62% for experimental studies and 20/28, 71% for design), consistent with controlled laboratory conditions or design experiments that isolate these dyadic interactions. Perspective papers (n = 28) show a similar concentration on single-user scenarios (18/28, 63%), with only modest attention to many-human to one AI configurations (3/28, 11%) – even though these studies are not limited by empirical constraints. Even qualitative research (n = 11), which might be expected to explore social dynamics, concentrates on single-user scenarios (8/11, 73%).

This methodological bias toward dyadic interactions has consequences: the social, networked dimensions of sensemaking, such as information sharing within and across larger groups, coordination of distributed decisions, emergence of collective understanding, remain largely unexplored in empirical HCAI research, i.e., this gap is systematic across research methods, not confined to a particular methodological tradition.

What is missing are studies that analyse the broader societal implications, collective intelligence, and democratic processes whereby many humans work with many AI systems that impact information flows and decision-making. When many actors optimise their choices based on optimised input by AI algorithms, their collective behaviour can generate system-wide effects that no single actor anticipated or can control. If AI is implemented at scale and optimised decisions propagate through and even shape the networks of our interaction, what are the emergent effects on human sensemaking and decision-making over time? And how can these emergent effects in complex human–AI networks be understood and controlled? The principles that guide the interaction of one human

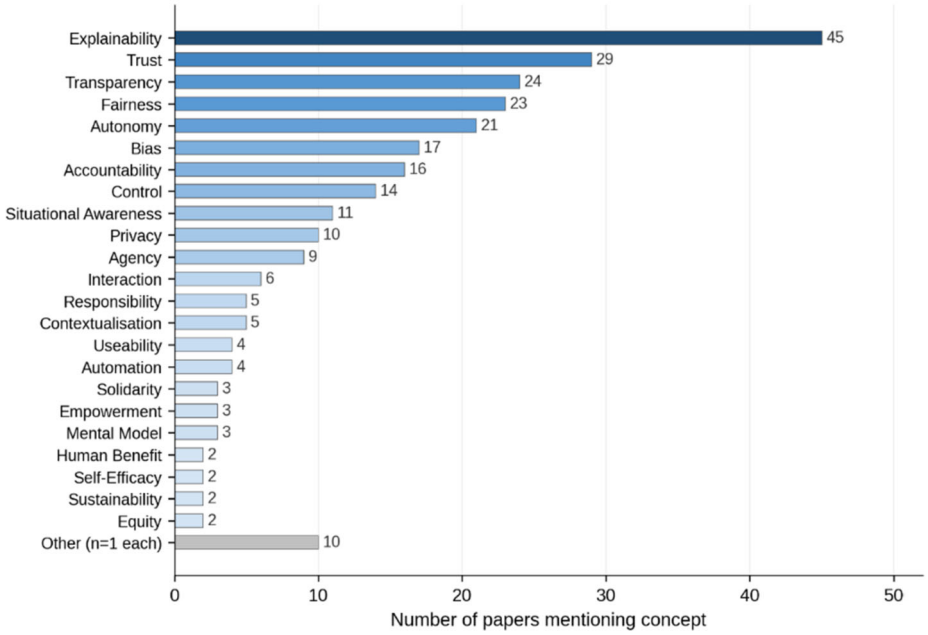


**Figure 7** Frequency of concepts in HCAI papers. Showing concepts with at least three mentions across the 101 papers. Explainability dominates the discourse ($n = 45$, 45%), followed by trust ($n = 29$, 29%), transparency ($n = 24$, 24%), and fairness ($n = 23$, 23%). Principles related to human oversight receive comparatively limited attention: control ($n = 14$, 14%), agency ($n = 9$, 9%), or solidarity and empowerment (each n = 3, 3%)

with one AI or one AI with a small group of humans may not suffice for analysing the dynamics of complex networks in which many humans are supported by many machines. These dynamics raise questions about automation, coordination, and the preservation of meaningful human control.

### 4.3 What HCAI optimises for

Decisions are inherently linked to what we value. AI that supports human decision-making must therefore also grapple with fundamental questions about which values and principles should guide their design and operation. As outlined in the methods section, the HCAI literature blurs the boundary between normative values and technical requirements. Figure 7 shows the concepts currently used as design principles and guiding values in the HCAI literature. Clearly, current HCAI research emphasises explainability (n = 45) (Sun [117]), trust (n = 29) (Liao & Sundar [73]), transparency (n = 24) (Kunar et al. [67]), and fairness (n = 23) (Kattnig et al. [62]), while other concepts ranging from equity (n = 2) (Akula & Garibay [4]) to solidarity (n = 3) (Sigfrids et al. [109]), humanity (n = 1, 'Other') (Comes [17]), or safety (n = 1, 'Other') (Zhang et al. [138]) receive far less attention, even though they are key to human decision-making, especially in poly-crises.

Given the analytical nature of AI research, it may not be surprising that also HCAI research prioritises concepts that can be easily measured or optimised for, such as explainability, fairness or trust. This creates a fundamental problem: explainability metrics become substitutes for accountability; fairness for justice; and trust replaces meaningful human oversight. The result is systems optimised for measurable proxies rather than the deeper values or principles they are supposed to represent. By focusing on what can be

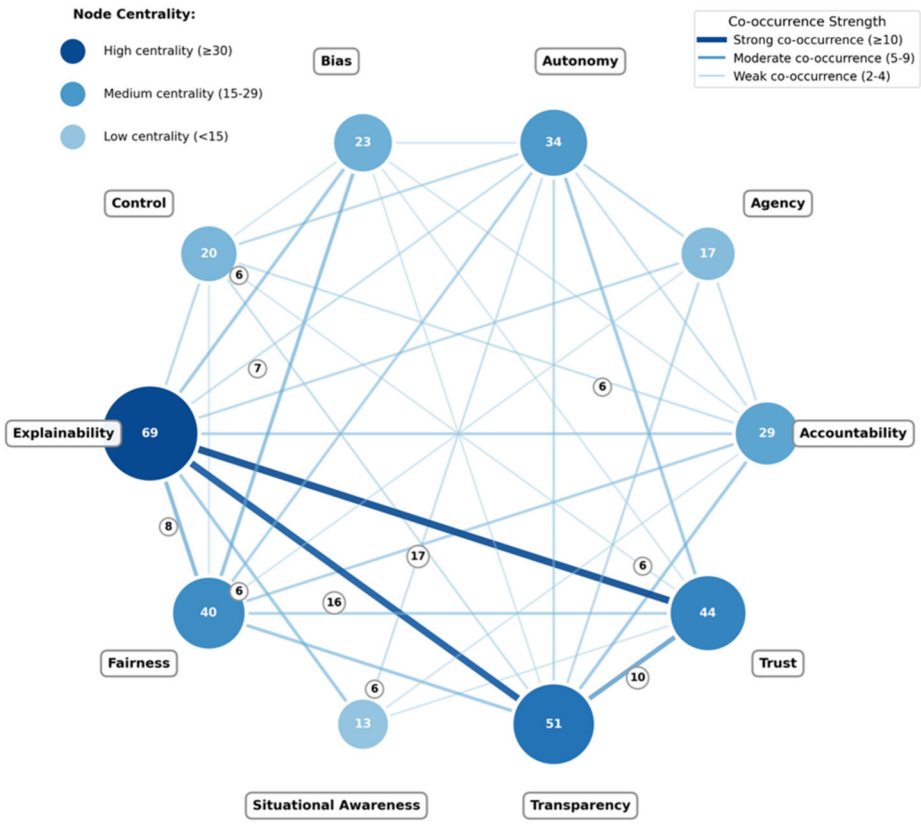


**Figure 8** Co-occurrence Network of HCAI Principles. Node values show centrality scores (frequency + number of unique co-occurring principles). Node size and colour reflect centrality, with darker/larger nodes indicating principles that are frequently mentioned and well-connected to other principles. Edge thickness indicates co-occurrence strength; numbers on edges ($\geq 6$) show how many papers discuss both principles together. Line color intensity corresponds to co-occurrence frequency, (darker = stronger)

optimised for, the field of HCAI may therefore inadvertently reproduce the very reductionism it seeks to address.

Moving from individual concepts to patterns of co-occurrence, Fig. 8 visualises the co-occurrence network of HCAI principles, revealing which principles dominate the discourse and how they cluster. Node centrality scores combine how often a principle appears (frequency) and how many other principles it co-occurs with (connectivity). From this analysis, explainability emerges as the most central principle (centrality: 69), appearing in 45 papers all other principles. Trust (44) and transparency (51) form a strongly connected cluster with explainability, with the explainability-trust pairing appearing in 17 papers, the strongest co-occurrence in the network. This cluster reflects the dominant framing of HCAI around AI transparency and user confidence. In contrast, principles related to human agency and control occupy peripheral positions with lower centrality scores for agency (17), control (20), and autonomy (34). These control-related principles rarely co-occur despite their conceptual relatedness, suggesting fragmentation in how the literature addresses meaningful human control. This fragmentation is particularly evident in the absence of strong connections between agency, control, and autonomy. Surprisingly, agency, autonomy and control are less frequently associated with fairness, accountability

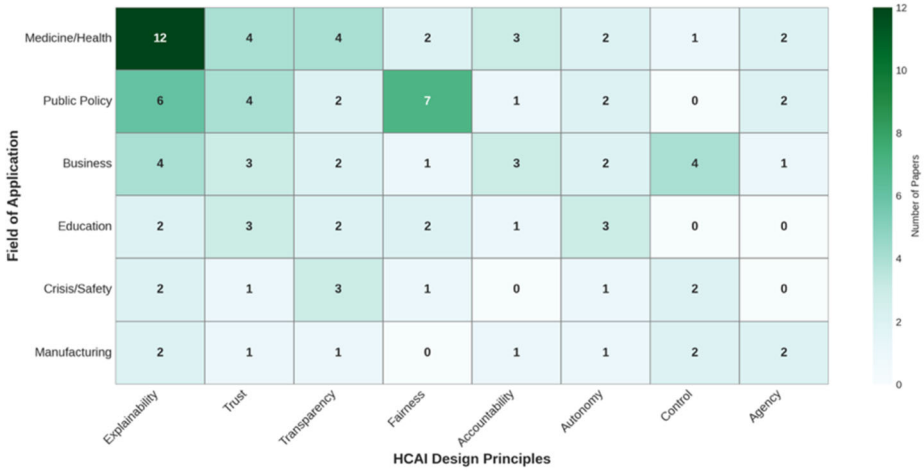


**Figure 9** Heatmap of HCAI Principles across domains of application. Medicine and health research ($n = 20$) prioritizes explainability (60%), trust and transparency (*each*20%), while public policy applications ($n = 12$) emphasize fairness (58%) alongside explainability (50%). Business applications ($n = 11$) show balanced attention across control (36%), explainability (36%), and accountability (27%). The diverse patterns reveal that HCAI principles are treated as domain-specific requirements rather than as a universal principle

and privacy, suggesting a research gap in how we balance the need for efficient, rapid and automated decisions with human oversight and value deliberation.

Analysing three-way combinations reveals concentrated clustering: the combination of explainability-trust-transparency appears in eight papers, representing a cluster concerned with the implications algorithmic transparency and understanding. Fairness-accountability-transparency, aligned with democratic governance, appears in only two papers. Most strikingly, a combination of agency-control-autonomy, concepts central to meaningful human oversight, never co-occurs in any single paper. This fragmentation intensifies at higher orders: examining four-way combinations, only one combination (explainability-fairness-transparency-trust) appears in more than one paper (n = 2). The absence of integrated combinations around human oversight reveals that HCAI research treats control, agency, and autonomy as decomposable concerns rather than aiming to understand how we can preserve meaningful human agency and control when AI systems shape decision architectures.

HCAI principles are domain specific, i.e., different application domains emphasise different HCAI principles, reflecting divergent concerns per domain or regulatory pressures. Figure 9 shows the distribution of HCAI principles across major application fields. Medicine and Health research (n = 20) emphasises explainability (12/20, 60%), transparency and trust (each 4/20, 20%), consistent with decision-making contexts where understanding the rationale of a decision is important for acceptance. However, fairness receives limited attention (2/20, 10%) despite well-documented health disparities. In contrast, public policy research (n = 12) prioritizes fairness (7/12, 58%) alongside explainability (6/12, 50%), reflecting possible concerns about algorithmic accountability in public service delivery. Yet accountability itself, despite being central to democratic governance, appears in only one paper (1/12, 8%), suggesting a gap between normative commitments and empirical research. For the business sector, the importance of control (4/11, 36%), ex-

plainability (4/11, 36%), and accountability (3/11, 27%) are balanced, indicating concerns about oversight and performance attribution

These domain-specific patterns and differences show that HCAI principles are treated as domain requirements rather than as fundamental principles of human-AI interaction. While this may be in line with calls for contextualised AI, especially in explainability (Liao et al. [74]), this fragmentation suggests that current HCAI research has not yet developed an understanding of which AI principles are universal or context-dependent; and related, there is no universal understanding of how AI shapes sensemaking across contexts.

In sum, this scoping review shows that current HCAI focuses on dyadic interactions, measurable principles, and task distribution between humans and machines. First, a focus on specific domains leads to principle fragmentation: different fields emphasise different principles based rather than developing integrated frameworks or an understanding of what are universal versus domain-specific principles. Second, methodological preferences for controlled, dyadic settings systematically neglect the networked, collective dynamics central to sensemaking in complex systems, even though in principle qualitative, perspective or large-scale survey paper should be able to capture these aspects. Third, the absence of consistent combinations of HCAI principles, particularly around agency, control, and autonomy, reveals that HCAI research itself is fragmented. In other words, AI systems can be 'explainable' to individual users, 'fair' in specific transactions, and offer 'control' at the interface level, yet still reshape how groups share information, how networks coordinate decisions, and how communities construct collective meaning, precisely the dynamics central to sensemaking in complex systems. These findings motivate the shift toward Sensemaking AI articulated in Sect. 5: addressing how AI participates in collective meaning-making requires moving beyond field-specific principles, individual-level analysis, and fragmented concepts toward integrated frameworks that recognize AI as an actor within evolving socio-technical networks.

## 5 Discussion: from human-centred to sensemaking AI

The results of the scoping review show a fundamental tension: while HCAI research has made important advances in improving individual-level explainability, fairness, or trust, it has yet to grapple with what happens when these optimised interactions scale to complex, networked environments where many humans work with many AI systems. In such environments, optimisation reshapes the informational and social conditions under which actors interpret, coordinate and decide – the terrain of sensemaking itself. This creates an *optimisation paradox*: the more we optimise individual human-AI interactions, the less equipped we may become to handle the emergent, collective challenges that define complex socio-technical systems. This paradox motivates a reorientation from model-centric improvements toward AI that sustains collective meaning-making in evolving human–AI networks.

This paradox be comes particularly visible in poly-crises, which are complex, decentralised, fraught with dilemmas and marked by '*time running out*' (Comes [17]; Levin et al. [71]). This combination fundamentally changes human sensemaking and decision-making behaviour, thereby also altering the evolving human-AI-interactions.

HCAI research predominantly frames interactions as generic, not contextualised, as dyadic, not networked and as static, not dynamically evolving. The networks involved in

crises, however, are inherently dynamic and uncertain, demanding flexible and context-sensitive AI systems (Jennings et al. [58]). Similarly, while other fields such as AI for Collective Intelligence (Gupta & Woolley [45]; Riedl & De Cremer [97]) or Human-Cyber-Physical Systems (Zhou et al. [139]) analyse the impact of AI on social-technical system embrace the idea of networks and larger groups, they assume stable tasks and roles. In contrast, Sensemaking AI addresses the uncertainty and ambiguity central to complex problems (Comes, Van Wassenhove, et al. [19]; Helbing [48]): situations where values, objectives, problem definitions, and roles emerge through collective interpretation and deliberation in dynamically evolving socio-technical networks. Building on these observations, I propose three shifts: Sensemaking-aware automation, collective agency for network-level control, and value-aware sensemaking. Together, these shifts provide a research and design agenda for Sensemaking AI.

### 5.1 Sensemaking-aware automation

It is tempting to assume that machines can optimise our meaning-making—make it faster, more efficient, less biased. But automation shapes the trajectories of interpretation by narrowing how problems are framed. Although the review shows the increasing recognition of human-AI-teaming in situational awareness (Endsley [33]; Gajos & Mamykina [41]) or developing shared mental models (Hoch et al. [52]; Thompson [120]), most research focuses on computer-supported information sharing or decision-making. Sensemaking is treated as an individual cognitive process that can be supported through better information presentation, rather than recognising it as a fundamentally *collective* process of meaning construction that emerges through social interaction and network dynamics. In addition, the AI is seen as a (neutral) tool that mediates human-human interaction (Heyndels [51]), rather than viewing AI as an *actor* that optimises information flows and thereby shapes sensemaking and decision-making. As humans rely routinely on AI, however, the relation between a human 'operator' and an AI tool supporting the human has become blurry.

Automation does not just take over individual tasks or decisions. Rather, automation reshapes how we understand what matters by impacting the process of meaning-making, through which we individually and collectively define and understand our choices. Via sensemaking trajectories, initial perceptions – as moderated or provided by an AI - can become deeply engrained (Comes, Van de Walle, et al. [18]), leading to path-dependencies that remain hidden within the current task-based frameworks. As AI systems continuously filter, prioritise, and present information, they do not merely 'support' decision-making but shape human cognition, reinforcing certain narratives while marginalising others. In essence, sensemaking depends on *ambiguity*, the interpretive flexibility that enables creative reframing when understanding evolves (Weick [134]). The question becomes not whether to automate information processing or decision-making, but how to ensure such automation supports rather than constrains the interpretive flexibility that enables creative reframing. A way ahead in achieving interpretive flexibility may be graceful degradation (Ploeg et al. [94]) – designed fallback from higher levels of automation that force systems to slow, expose provenance, invite dissent, and hand decision authority to people if needed. Interpretive flexibility, however, is not an individual cognitive capacity that can be preserved through dyadic human-AI interactions. Sensemaking is fundamentally social—it emerges through collective interaction, shared interpretation, and distributed

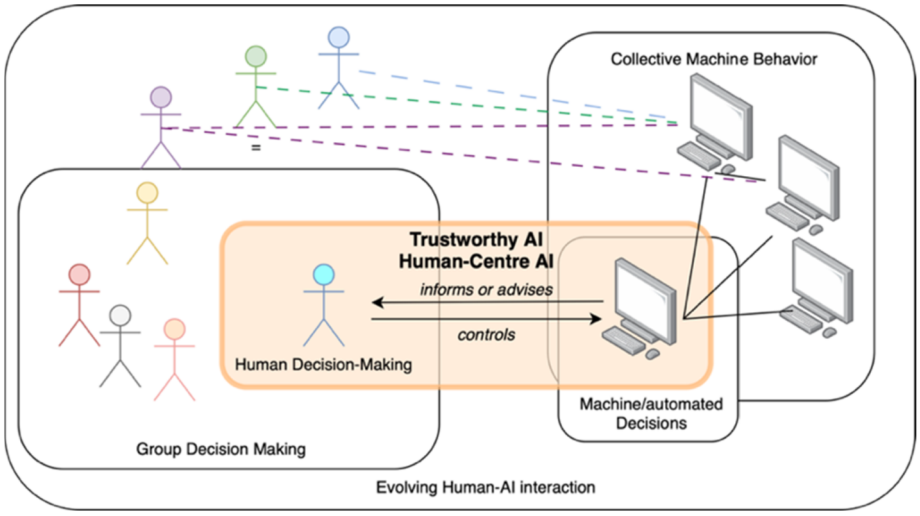


**Figure 10** The layered interactions of humans and machines – from Human-Centred AI to Sensemaking AI. Yellow box in the middle showcases the dominant paradigm of HCAI Research and Trustworthy AI. I argue to extend the realm to Sensemaking AI, also using principles of Group Decision Making and Collective Machine Behaviour

meaning-making across networks of actors (Weick [133]). Yet current HCAI research largely overlooks this social dimension.

Especially the role of AI in identity construction, by which identity is continuously shaped through social interaction and feedback (Weick [133]), has not received attention within HCAI yet. AI, as an actor, influences how humans construct their professional and social identities by mediating access to information, influencing real or perceived agency, and reinforcing or challenging organisational and societal norms, as is for instance shown by the impact of ChatGPT on the identity of researchers (Hu et al. [55]). Similarly, AI alignment processes that rely on learning from human feedback (Dahlgren Lindström et al. [25]) may inadvertently shape user identities by reinforcing certain value expressions while marginalizing others, raising questions about whose values are being aligned to and how this shapes collective identity construction over time. Especially with the increasing personalisation of generative AI, there is a risk that identities become shaped by an algorithm that amplifies biases and leads to echo chambers. When AI systems optimise for engagement, productivity, or efficiency, they reshape the very questions humans ask about meaning and purpose. In addition, there is a growing concern about the risk of de-skilling, where the reliance on AI leads to an erosion of critical thinking (Sellen & Horvitz [103]). What is more, complex problems are often characterised by moral dilemmas. Yet off-loading morally challenging decisions to a machine may lead to *moral de-skilling* (Vallor [124]). As such, the impact of AI on perceived or real responsibility for others, and on the evolution of human competences and skills are a concern that must be addressed in sensemaking aware automation.

To address this gap, Sensemaking AI can draw on different bodies of literature as summarised in Fig. 10. First, there is wealth of research on group decision-making (Hollingshead et al. [54]), information sharing and sensemaking (Stasser & Titus [113]) (left box in Fig. 10; see also Fig. 2) dedicated to how teams and groups differ from individuals, which is largely neglected in the HCAI literature. This gap becomes evident when we consider

our finding that only 22 (9 + 13) out of 101 papers address situations where AI systems interact with multiple humans simultaneously. At the other end of the spectrum, research on collective machine behaviour and multi-AI coordination (right side in Fig. 10) focuses on coordinating different artificial agents (Stone et al. [115]). While 13 papers in the review discuss autonomous systems, none of them focuses how the coordination of these systems can or should be designed to benefit humans. Here, the role of trust, loyalty, or cognitive and behavioural factors that are important in human interactions and group decision-making are discarded; networks and groups are formed based on optimal skillsets or available resources.

I argue that Sensemaking AI should address the reality of complex networks where multiple types of interactions occur simultaneously: humans engaging in collective sensemaking with each other, humans interacting with multiple AI systems that shape their information environments and thereby their decisions, and AI systems that coordinate or influence each other. This implies that theories of Sensemaking AI recognise that all these interactions constitute a single, complex system where human sensemaking, AI mediation, and algorithmic coordination are fundamentally intertwined and mutually constitutive. This requires integrating theories on group decisions and sensemaking into HCAI research and incorporating task distribution and information prioritisation protocols from collective machine behaviour research.

Building on the integrated theoretical framework and recognising that AI fundamentally influences social networks, shared meaning-making and values, we need to then focus on Sensemaking AI as a design principle and ask: which sensemaking, coordination processes and decisions do we want to or need to optimise or automate, and why? Answering this question, especially given the urgency and moral dilemmas pertaining to poly-crises, requires addressing research questions such as: How does AI impact shared identity construction and how can collective moral de-skilling be avoided? How do optimisation algorithms impact collective sensemaking trajectories and thereby shape the interpretive flexibility that is crucial for sensemaking?

Answering these questions requires expanding the current theoretical foundations and conduct interdisciplinary research that focuses on large-scale and longitudinal studies on network dynamics as the central focus. Here, longitudinal empirical studies with foundations in group decision making (e.g., information pooling, shared mental models) can be combined with insights from machine behaviour (e.g., prioritisation protocols) and computational models from complexity science to capture the interplay between automation and evolving sensemaking. Experimental designs that measure interpretive flexibility, identity construction and (moral) deskilling before and after the use of an AI in different constellations of groups and teams can create new insights into sensemaking trajectories; based on these empirical insight, agent-based models (Nespeca et al. [88]) can simulate information sharing and sensemaking dynamics to explore path-dependencies and conditions or tipping points that lead to the erosion of interpretive flexibility in networks.

### 5.2 Collective agency for network-level control

The discussion about automation is inherently connected to questions of control, autonomy and agency. However, this review shows limited attention for control mechanisms beyond individual human-AI interactions, with concepts like 'control' (n = 12) and 'agency' (n = 10) receiving far less attention than principles like explainability (n = 45), see Fig. 8.

Even 'autonomy' (n = 21) is most often discussed in the context of trust (6/21) and fairness (6/21) rather than through the lens of control (4/21), see Fig. 9.

Current principled approaches to control and accountability overlook the complexity arising from the many diverse interactions of humans and machines. For instance, the OECD guidelines specify that "*AI actors should be accountable for the proper functioning of AI systems*" (OECD [89]). But what if it is precisely the 'proper functioning' that leads to undesired consequences or harmful cascading effects? When optimisation decisions propagate through networks of human and AI actors, the "problem of many hands" and subsequent responsibility gaps occur (Matthias [82]). This problem cannot simply be solved by distributing responsibility (Coeckelbergh [15]) since it is not clear who bears responsibility when properly functioning optimisation algorithms produce undesirable outcomes at systems level. Preserving human control in such dynamic systems thus requires *temporal reflexivity* (see Sect. 2), i.e., the ability to recognise when optimisation drifts away from the original purpose and to intervene in time to change the unfolding trajectories when needed.

Moreover, research has shown that decisions shape physical, informational, and social networks, which in turn influence the information accessible (Comes, Van de Walle, et al. [18]) to human actors and AI agents. When algorithms optimise traffic and energy flows, financial markets, or entire smart cities, they create optimised environments, in which human choices are increasingly constrained by algorithmic assumptions about what should be optimised, and how. In these contexts, traditional concepts of control—rooted in task allocation and oversight—become inadequate since the challenge is not only controlling what machines *do*, but preserving spaces for human interaction and meaning-making within emergent, decentralised networks. As such, the question of human control becomes: how can human agency be preserved when optimisation algorithms and AI increasingly dominate information flows and decision architectures? How do decision-information feedback loops influence the long-term evolution of control structures in human-AI networks across spatial and temporal scales?

Addressing control in emergent human-AI networks requires moving beyond traditional oversight models toward networked agency. Control theory and cybernetics, originally developed by Wiener [135], provides the framework for understanding adaptive regulation: control theory provides a framework for modelling decision loops where human and AI agents dynamically adjust their actions based on new information, constraints, and goals. Cybernetics stresses the need for self-correcting incentives and governance. As such, cybernetics has also been suggested as a way to coordinate decentralised AI networks in autonomic computing (De Wolf & Holvoet [28]) and more recently as a governance principle for humans and technology (Zwitter [140]). Rather than centralised monitoring, cybernetic approaches enable *network-level self-regulation* where control emerges through distributed feedback loops and adaptive responses to changing conditions, e.g., via meta-signals on uncertainty and impact, circuit breakers for cascading automations, auditability of decision–information feedback loops. This shift reframes control from actions or outcomes to designing self-correcting conditions, under which collective agency can emerge and evolve with human-AI networks.

### 5.3 Value-aware Sensemaking AI: processes and boundaries

Value-aware Sensemaking AI refers to AI that makes value assumptions explicit and revisable. As such, value-aware Sensemaking AI needs to distinguish between process princi-

ples (how values are formed, contested and revised) and content principles. Such systems make value assumptions explicit and revisable, facilitating how values are surfaced, balanced and formalised while preserving space for disagreement if trade-offs violate moral boundaries.

The HCAI literature recognises that AI systems need to be designed to "*understand humans*" including the norms and values that govern our actions (Riedl [98]). Even though Shneiderman [107] proposed HCAI as a design *process*, this review shows '*human-centred*' has largely become a synonym to explainable, fair, accountable, transparent and trusted AI systems, see Fig. 8. These principles are often treated as generic optimisable requirements that systems can be built from irrespective of the context. Maybe not surprisingly, a similar view is presented by the various guidelines, standards and regulatory frameworks for the design and use of AI. The UNESCO recommendations on the Ethics of AI (UNESCO [123]), the OECD Recommendation of the Council on Artificial Intelligence (OECD [89]), the European Commission's recommendations by the High-Level Expert Group on AI (EC [30]) the IEEE standards for Ethically Aligned Design of Autonomous and Intelligent Systems (IEEE [56]), and the EU AI Act establish important foundations around transparency, accountability, trust and fairness. However, they all operate under the assumption that values can be translated into generic, stable, measurable principles. This suggests a tendency to optimise for what can be measured while neglecting values that may be essential *because* they resist quantification.

This challenge is further compounded by how AI alignment is conceptualised. AI alignment, the effort to ensure AI systems behave in line with human intentions and values, has become a central concern in AI safety (Ji et al. [59]). The idea is that by ensuring that human goals are matched, unintended outcomes or AI failures such as reward hacking can be avoided (D'Amato [26]). However, AI alignment approaches typically assume that human values can be specified as stable objectives to which AI systems can and should be aligned, either through explicit programming, learning from human feedback or aggregating human preferences via social choice theory (Conitzer et al. [20]). However, these approaches presuppose what Sensemaking AI questions: that we know what values to align to *before* engaging with complex problems.

In addition, current HCAI principles assumes that higher level principles can be achieved simultaneously. Yet, there are inherent conflicts across principles or what society values, and these conflicts cannot always be reconciled, e.g., when climate justice collides with economic stability, transparency with privacy, or control with personal freedom. There is an expanding literature that highlights that humans refuse making such trade-offs because they are seen as morally problematic, or taboo (Chorus et al. [14]; Tetlock [119]). Formalising such trade-offs in an optimisation then risks treating them as commensurable, thereby eroding their role as ethical boundaries.

The underlying challenge is: AI principles are viewed in separation from the contextualised objectives, preferences, attitudes, or emotions that drive human sensemaking and decision-making (van Berkel et al. [126]), as well as from the consequences that occur if AI is used at scale. Based on sensemaking theory, I argue that the meaning of key principles depends on continuous collective (re-)interpretation. I do not advocate for abandoning principled approaches, but rather for distinguishing between process principles that support collective meaning-making and democratic deliberation from content principles that may predetermine its outcomes. This shift requires integrating approaches that recognise

value formation as an *output* of human-AI interaction, rather than as an input as is done e.g., in AI alignment. Social choice ethics for AI design (Baum [8]) provides a framework for this, emphasising questions of standing (*who* participates in value construction?), measurement (*how* are diverse perspectives translated into system design?), and aggregation (*how* do we coordinate across potentially conflicting value systems?).

This process-centered approach creates an important challenge: how to translate the outcomes of collective deliberation into formal specifications without undermining the integrity of the process? To ensure that the results of collective value formation can guide AI development, outcomes need to be linked to formal decision theoretical frameworks that translate values into objective functions and operational trade-offs. Research is needed to formalise abstract goals such as equity (Coleman et al. [16]; Holguin-Veras et al. [53]) and to explore the dynamic nature and structure of trade-offs for instance for intangible or sacred goods via taboo trade-offs (Daw et al. [27]; Lu et al. [76]). At the same time, democratic deliberation may identify domains, values or decisions that cannot be translated, and where preserving ambiguity, maintaining human judgment, and sustaining ongoing deliberation is more important than efficiency. This leads to the question: how can AI systems recognise and respect the boundaries of their own applicability as determined via deliberation? By addressing these challenges, research on Sensemaking AI can combine the cognitive, behavioural, social and ethical elements needed to move towards AI that supports rather than constrains collective meaning-making.

Taken together, the three shifts towards Sensemaking AI repositions AI from a tool that optimises towards generic objectives in dyadic relations to an actor that shapes collective cognition in networks. AI and humans, together, are tasked with sustaining interpretive flexibility, networked control and value-awareness at scale.

## 6 Conclusion

Human-Centred AI (HCAI) has been put forward as a paradigm to design AI the supports humans by advocating for design principles ensuring that AI is explainable, fair, transparent and trustworthy. Yet, this scoping review shows that these principles are largely operationalised for dyadic interactions where one human works with one AI, and that the focus is on a relatively narrow set of values that can be readily operationalised. Even though I acknowledge that this review provides only a snapshot, and that the field of (HC)AI is rapidly evolving, this framing is too narrow for the complex and often time-compressed decisions that we are facing today. The ubiquity of AI, and the optimisation of information sharing and processing, reconfigures the informational and social conditions under which humans interpret situations and decide. AI reshapes the way we make sense of our environment.

Against this backdrop, this paper advocates for a paradigm shift towards Sensemaking AI: AI that supports collective meaning-making in evolving human-AI networks. Conceptually, the paper synthesises sensemaking and decision theory with literatures on coordination and machine behaviour to characterise AI as an actor within socio-technical systems. Via a scoping review of the HCAI literature, this paper highlights that current research focuses on individual support and generic, measurable principles. Gaps persist in our understanding of how (human-centred) AI impacts and reshapes sensemaking and decision-making over time. Together, these strands motivate three interconnected directions for Sensemaking AI research and design:

- **Sensemaking-aware automation:** AI shapes sensemaking trajectories, reinforcing certain narratives while marginalising others. Research needs to expand beyond dyadic human-AI interactions and integrate group decision and collective machine behaviour theories to understand how automation impacts collective sensemaking in dynamic networks where many humans work with many algorithms. Future research questions include: How does AI-driven automation influence identity construction, (moral) de-skilling and collective meaning-making over time? What mechanisms can mitigate path dependencies, and how to preserve the interpretive flexibility essential for creative reinterpretation in networked systems.
- **Collective agency for networked control:** in complex systems, oversight of individual components can never ensure control of the whole system. Therefore, AI understood as an actor in complex networks poses a challenge for conceptualising and maintaining human control and oversight, especially since over time optimisation cascades create decision-information feedback loops. Sensemaking AI reframes control as networked agency. Research needs to analyse how to preserve collective agency when optimisation logic increasingly dominates decision architectures. Drawing on control theory and cybernetics, this requires designing conditions that enable network-level self-regulation through distributed feedback loops and adaptive responses to changing conditions.
- **Value-aware Sensemaking:** Current principled AI frameworks lack the adaptability required for complex, dynamic decisions. Recognising that some values—such as dignity, justice, and humanity—resist translation into optimisable metrics and cannot be traded off, Sensemaking AI distinguishes process principles, which support ongoing democratic deliberation and contextualisation, from content principles that risk pre-empting it. Methodologically, this calls for pipelines that (i) enable participatory formation and revision of objectives, (ii) translate deliberative outcomes into formal decision models *where appropriate*, and (iii) specify boundaries of applicability where automated optimisation should defer to human judgment and sustain ambiguity. The aim is not to abandon principles, but to embed them in processes that keep values contestable and revisable at scale.

The shift towards Sensemaking AI has actionable implications for research and practice: the focus of AI studies has to shift from individual decision-makers to studies that acknowledge AI as embedded in evolving social-technical networks. This requires a shift of methods towards integrating longitudinal, large-scale empirical studies with methods from complexity science to trace how information sharing, analysis, explanation and automation affect interpretive flexibility, sensemaking and coordination over time. AI design should integrate and test mechanisms to ensure a diversity of inputs, allow for surfacing dissent rather than focusing on convergence, and allow to shift from automation to human deliberation when needed. AI governance needs to move towards architectures that integrate feedback mechanisms and incentives, rather than promoting static principles.

The contribution of this paper lies in challenging the imperative within current Human-Centred AI literature to optimise AI systems to become more transparent or fair. Instead, Sensemaking AI is proposed as a concrete alternative that recognises AI as an actor within complex socio-technical systems that shapes collective meaning making and decision architectures. As such, Sensemaking AI needs to be designed to sustain the interpretive, social, and ethical capacities on which sensemaking depends.

**Abbreviations**
AI, Artificial Intelligence; EC, European Commission; HCAI, Human-Centred Artificial Intelligence; LLM, Large Language Model; NLP, Natural Language Processing; OECD, Organisation for Economic Co-operation and Development; PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

## Supplementary information

**Supplementary information** accompanies this paper at https://doi.org/10.1140/epjds/s13688-026-00634-5.

**Additional file 1.** (DOCX 32 kB)

**Acknowledgements**
Not applicable.

**Author contributions**
Tina Comes is the sole author of this manuscript.

**Funding information**
No funding was received to assist with the preparation of this manuscript.

**Data availability**
No datasets were generated or analysed during the current study.

## Declarations

**Clinical trial number**
Not applicable as this is not a clinical study.

**Ethics approval and consent to participate**
Not applicable since there was no participation of human subjects in the study.

**Consent for publication**
Not applicable since this was the work of the author.

**Competing interests**
The authors declare no competing interests.

Received: 26 August 2025 Accepted: 24 February 2026 Published online: 19 March 2026

**References**

1. Abbass HA (2019) Social integration of artificial intelligence: functions, automation allocation logic and human-autonomy trust. Cogn Comput 11(2):159–171. https://doi.org/10.1007/s12559-018-9619-0
2. Acharya DB, Kuppan K, Divya B (2025) Agentic AI: autonomous intelligence for complex goals—a comprehensive survey. IEEE Access 13:18912–18936. https://doi.org/10.1109/ACCESS.2025.3532853
3. Akata Z, Balliet D, de Rijke M, Dignum F, Dignum V, Eiben G, Fokkens A, Grossi D, Hindriks K, Hoos H, Hung H, Jonker C, Monz C, Neerincx M, Oliehoek F, Prakken H, Schlobach S, van der Gaag L, van Harmelen F, et al (2020) A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. Computer 53(8):18–28. https://doi.org/10.1109/MC.2020.2996587
4. Akula R, Garibay I (2021) Ethical AI for social good. In: HCI international 2021 - late breaking papers: multimodality, eXtended reality, and artificial intelligence, Cham
5. Alix C, Lafond D, Mattioli J, Heer JD, Chattington M, Robic PO (2021) Empowering adaptive human autonomy collaboration with artificial intelligence. In: 2021 16th international conference of System of Systems Engineering (SoSE)
6. Auernhammer J (2020) Human-centered AI: the role of Human-centered Design Research in the development of AI
7. Bansal G, Wu T, Zhou J, Fok R, Nushi B, Kamar E, Ribeiro MT, Weld D (2021) Does the whole exceed its parts? The effect of ai explanations on complementary team performance. CHI, New York
8. Baum SD (2020) Social choice ethics in artificial intelligence. AI Soc 35(1):165–176. https://doi.org/10.1007/s00146-017-0760-1
9. Bingley WJ, Haslam SA, Steffens NK, Gillespie N, Worthy P, Curtis C, Lockey S, Bialkowski A, Ko RKL, Wiles J (2023) Enlarging the model of the human at the heart of human-centered AI: a social self-determination model of AI system impact. New Ideas Psychol 70. https://doi.org/10.1016/j.newideapsych.2023.101025
10. Buçinca Z (2024) Optimizing decision-maker's intrinsic motivation for effective human-AI decision-making. Extended abstracts of the CHI conference on human factors in computing systems. https://doi.org/10.1145/3613905.3638179
11. Capel T, 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, Hamburg, Germany. https://doi-org.tudelft.idm.oclc.org/10.1145/3544548.3580959
12. Cavalcante Siebert L, Lupetti ML, Aizenberg E, Beckers N, Zgonnikov A, Veluwenkamp H, Abbink D, Giaccardi E, Houben G-J, Jonker CM (2023) Meaningful human control: actionable properties for AI system development. AI Ethics 3(1):241–255

13. Chaudhry MA, Cukurova M, Luckin R (2022) A transparency index framework for AI in education. In: Artificial intelligence in education. Posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners' and doctoral consortium, Cham
14. Chorus CG, Pudāne B, Mouter N, Campbell D (2018) Taboo trade-off aversion: a discrete choice model and empirical analysis. J Choice Model 27:37–49. https://doi.org/10.1016/j.jocm.2017.09.002
15. Coeckelbergh M (2020) Artificial intelligence, responsibility attribution, and a relational justification of explainability. Sci Eng Ethics 26(4):2051–2068
16. Coleman N, Li X, Comes T, Mostafavi A (2024) Weaving equity into infrastructure resilience research: a decadal review and future directions. npj Nat Hazards 1(1):25. https://doi.org/10.1038/s44304-024-00022-x
17. Comes T (2024) AI for crisis decisions. Ethics Inf Technol 26(1):12. https://doi.org/10.1007/s10676-024-09750-0
18. Comes T, Van de Walle B, Van Wassenhove L (2020) The coordination-information bubble in humanitarian response: theoretical foundations and empirical investigations. Prod Oper Manag 29(11):2484–2507
19. Comes T, Van Wassenhove L, Van de Walle BA (2020) The coordination-information bubble in humanitarian response: theoretical foundations and empirical investigations. In: Production and operations management, online first. https://doi.org/10.1111/poms.13236
20. Conitzer V, Freedman R, Heitzig J, Holliday WH, Jacobs BM, Lambert N, Mossé M, Pacuit E, Russell S, Schoelkopf H (2024) Social choice should guide ai alignment in dealing with diverse human feedback. arXiv preprint arXiv:2404.10271
21. Cooke N, Demir M, Huang L (2020) A framework for human-autonomy team research. In: HCII 2020, Copenhagen, Denmark
22. Coppi G, Moreno Jimenez R, Kyriazi S (2021) Explicability of humanitarian AI: a matter of principles. J Int Humanit Action 6(1):19
23. Council NR (1998) The future of air traffic control: human operators and automation. (C. Wickens, A. Mavor, R. Parasuraman, & J. McGee, Eds.). National Academies Press
24. Criscuolo C, Dolci T (2024) Exploring fairness interpretability with FairnessFriend: a chatbot solution. In: 2024 IEEE 40th International Conference on Data Engineering Workshops (ICDEW)
25. Dahlgren Lindström A, Methnani L, Krause L, Ericson P, de Rituerto de Troya ÍM, Coelho Mollo D, Dobbe R (2025) Helpful, harmless, honest? Sociotechnical limits of AI alignment and safety through reinforcement learning from human feedback: AD Lindström et al. Ethics Inf Technol 27(2):28
26. D'Amato K (2025) ChatGPT: towards AI subjectivity. AI Soc 40(3):1627–1641
27. Daw TM, Coulthard S, Cheung WWL, Brown K, Abunge C, Galafassi D, Peterson GD, McClanahan TR, Omukoto JO, Munyi L (2015) Evaluating taboo trade-offs in ecosystems services and human well-being. Proc Natl Acad Sci USA 112(22):6949–6954. https://doi.org/10.1073/pnas.1414900112
28. De Wolf T, Holvoet T (2003) Towards autonomic computing: agent-based modelling, dynamical systems analysis, and decentralised control
29. Duan X, Pei B, Ambrose GA, Hershkovitz A, Cheng Y, Wang C (2024) Towards transparent and trustworthy prediction of student learning achievement by including instructors as co-designers: a case study. Educ Inf Technol 29(3):3075–3096. https://doi.org/10.1007/s10639-023-11954-8
30. EC (2019) Ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
31. Elahi H, Castiglione A, Wang G, Geman O (2021) A human-centered artificial intelligence approach for privacy protection of elderly App users in smart cities. Neurocomputing 444:189–202. https://doi.org/10.1016/j.neucom.2020.06.149
32. Endsley MR (2017) From here to autonomy: lessons learned from human–automation research. Hum Factors 59(1):5–27
33. Endsley MR (2023) Supporting human-AI teams:transparency, explainability, and situation awareness. Comput Hum Behav 140, Article ID 107574. https://doi.org/10.1016/j.chb.2022.107574
34. Erlei A (2024) Understanding choice independence and error types in human-AI collaboration. In: Proceedings of the CHI conference on human factors in computing systems. https://doi.org/10.1145/3613904.3641946
35. Fan C, Zhang C, Yahja A, Mostafavi A (2021) Disaster city digital twin: a vision for integrating artificial and human intelligence for disaster management. Int J Inf Manag 56, Article ID 102049
36. Fitts PM (1951) Human engineering for an effective air-navigation and traffic-control system
37. Flathmann C, Schelble BG, Rosopa PJ, McNeese NJ, Mallick R, Madathil KC (2023) Examining the impact of varying levels of AI teammate influence on human-AI teams. Int J Hum-Comput Stud 177. https://doi.org/10.1016/j.ijhcs.2023.103061
38. Freire SK, Niforatos E, Wang C, Ruiz-Arenas S, Foosherian M, Wellsandt S, Bozzon A (2023) Lessons learned from designing and evaluating CLAICA: a continuously learning AI cognitive assistant. In: Proceedings of the 28th international conference on intelligent user interfaces, Sydney, NSW, Australia. https://doi-org.tudelft.idm.oclc.org/10.1145/3581641.3584042
39. French S (1986) Decision theory: an introduction of the mathematics of rationality. Ellis Harwood Limited
40. French S (2012) Cynefin, statistics and decision analysis. J Oper Res Soc. https://doi.org/10.1057/jors.2012.23
41. Gajos KZ, Mamykina L (2022) Do people engage cognitively with AI? Impact of AI assistance on incidental learning. In: Proceedings of the 27th international conference on intelligent user interfaces, Helsinki, Finland. https://doi-org.tudelft.idm.oclc.org/10.1145/3490099.3511138
42. Gil Y, Garijo D, Khider D, Knoblock CA, Ratnakar V, Osorio M, Vargas H, Pham M, Pujara J, Shbita B (2021) Artificial intelligence for modeling complex systems: taming the complexity of expert models to improve decision making. ACM Trans Interact Intell Syst 11(2):1–49
43. Gou J, Liang Q, Wang Z, Dabić M (2024) Affordances and constraints of automation and augmentation: lessons learned from development of a human-AI collaboration business simulation platform. J Glob Inf Manag 32(1):1–27. https://doi.org/10.4018/JGIM.357260
44. Gralla E, Goentzel J, Fine C (2016) Problem formulation and solution mechanisms: a behavioral study of humanitarian transportation planning. Prod Oper Manag 25(1):22–35. https://doi.org/10.1111/poms.12496

45. Gupta P, Woolley AW (2021) Articulating the role of artificial intelligence in collective intelligence: a transactive systems framework. Proc Hum Factors Ergon Soc Annu Meet 65(1):670–674. https://doi.org/10.1177/1071181321651354c
46. Hagemann V, Rieth M, Suresh A, Kirchner F (2023) Human-AI teams—Challenges for a team-centered AI at work. Front Artif Intell 6. https://doi.org/10.3389/frai.2023.1252897
47. He J, Piorkowski D, Muller MJ, Brimijoin K, Houde S, Weisz JD (2023) Understanding how task dimensions impact automation preferences with a conversational task assistant. CHI AutomationXP, Hamburg
48. Helbing D (2009) Managing complexity in socio-economic systems. Eur Rev 17:423–438
49. Helms Mills J, Thurlow A, Mills AJ (2010) Making sense of sensemaking: the critical sensemaking approach. Qual Res Organ Manage Int J 5(2):182–195. https://doi.org/10.1108/17465641011068857
50. Herrmann T, Pfeiffer S (2023) Keeping the organization in the loop: a socio-technical extension of human-centered artificial intelligence. AI Soc 38(4):1523–1542. https://doi.org/10.1007/s00146-022-01391-5
51. Heyndels S (2023) Technology and neutrality. Philos Technol 36(4):75
52. Hoch T, Heinzl B, Czech G, Khan M, Waibel P, Bachhofner S, Kiesling E, Moser B (2022) Teaming. AI: enabling human-AI teaming intelligence in manufacturing. In: I-ESA, Valencia, Spain
53. Holguin-Veras J, Perez N, Jaller M, Van Wassenhove LN, Aros-Vera F (2013) On the appropriate objective function for post-disaster humanitarian logistics models. J Oper Manag 31(5):262–280. https://doi.org/10.1016/j.jom.2013.06.002
54. Hollingshead AB, McGrath JE, O'Connor KM (1993) Group task performance and communication technology: a longitudinal study of computer-mediated versus face-to-face work groups. Small Group Res 24(3):307–333. https://doi.org/10.1177/1046496493243003
55. Hu H, Zhou Q, Hashim H (2025) Negotiating identity in the age of ChatGPT: non-native English researchers' experiences with AI-assisted academic writing. Humanit Soc Sci Commun 12(1):1–11
56. IEEE (2019) Ethically aligned design - a vision for prioritizing human well-being with autonomous and intelligent systems. In: Ethically aligned design - a vision for prioritizing human well-being with autonomous and intelligent systems. IEEE, Los Alamitos, pp 1–294
57. Jacob C, Kerrigan P, Bastos M (2025) The chat-chamber effect: trusting the AI hallucination. Big Data Soc 12(1), Article ID 20539517241306345
58. Jennings NR, Moreau L, Nicholson D, Ramchurn S, Roberts S, Rodden T, Rogers A (2014) Human-agent collectives. Commun ACM 57(12):80–88
59. Ji J, Qiu T, Chen B, Zhou J, Zhang B, Hong D, Lou H, Wang K, Duan Y, He Z, Vierling L, Zhang Z, Zeng F, Dai J, Pan X, Xu H, O'Gara A, Ng K, Tse B, et al (2025) AI alignment: a contemporary survey. ACM Comput Surv 58(5):132. https://doi.org/10.1145/3770749
60. Jin L, Boden A, Shajalal M (2022) Automated decision making systems in smart homes: a study on user engagement and design, vol 3154
61. Kahneman D, Tversky A (1979) Prospect theory: an analysis of decision under risk. Econometrica 47(2):263–291
62. Kattnig M, Angerschmid A, Reichel T, Kern R (2024) Assessing trustworthy AI: technical and legal perspectives of fairness in AI. Comput Law Secur Rev 55, Article ID 106053. https://doi.org/10.1016/j.clsr.2024.106053
63. Kiesler S, Sproull L (1992) Group decision making and communication technology. Organ Behav Hum Decis Process 52(1):96–123. https://doi.org/10.1016/0749-5978(92)90047-B
64. Kim JS, Kim M, Baek TH (2025) Enhancing user experience with a generative AI chatbot. Int J Hum-Comput Interact 41(1):651–663. https://doi.org/10.1080/10447318.2024.2311971
65. Kjærum A, Madsen BS (2025) Pushing the boundaries of anticipatory action using machine learning. Data Policy 7:e8
66. Klein G, Calderwood R, Clinton-Cirocco A (2010) Rapid decision making on the fire ground: the original study plus a postscript. J Cogn Eng Decis Mak 4(3):186–209. https://doi.org/10.1518/155534310X12844000801203
67. Kunar MA, Montana G, Watson DG (2024) Increasing transparency of computer-aided detection impairs decision-making in visual search. Psychon Bull Rev. https://doi.org/10.3758/s13423-024-02601-5
68. Lee MH, Siewiorek DP, Smailagic A, Bernardino A, Bermúdez i Badia S (2022) Towards efficient annotations for a human-ai collaborative, clinical decision support system: a case study on physical stroke rehabilitation assessment. IUI, Helsinki
69. Lee MK, Kusbit D, Kahng A, Kim JT, Yuan X, Chan A, See D, Noothigattu R, Lee S, Psomas A, Procaccia AD (2019) Webuildai: participatory framework for algorithmic governance. In: Proceedings of the ACM on human-computer interaction, vol 3. https://doi.org/10.1145/3359283
70. Lettieri N, Guarino A, Zaccagnino R, Malandrino D (2023) Keeping judges in the loop: a human–machine collaboration strategy against the blind spots of AI in criminal justice. Soft Comput 27(16):11275–11293. https://doi.org/10.1007/s00500-023-08604-z
71. Levin K, Cashore B, Bernstein S, Auld G (2012) Overcoming the tragedy of super wicked problems: constraining our future selves to ameliorate global climate change. Policy Sci 45(2):123–152
72. Li X, Zong Q, Cheng M (2024) The impact of medical explainable artificial intelligence on nurses' innovation behaviour: a structural equation modelling approach. J Nurs Manag 2024(1), Article ID 8885760. https://doi.org/10.1155/2024/8885760
73. Liao QV, Sundar SS (2022) Designing for responsible trust. In: AI systems: a communication perspective proceedings of the 2022 ACM conference on fairness, accountability, and transparency, Seoul, Republic of Korea. https://doi-org.tudelft.idm.oclc.org/10.1145/3531146.3533182
74. Liao QV, Zhang Y, Luss R, Doshi-Velez F, Dhurandhar A (2022) Connecting algorithmic research and usage contexts: a perspective of contextualized evaluation for explainable AI. Proc AAAI Conf Hum Comput Crowd 10(1):147–159. https://doi.org/10.1609/hcomp.v10i1.21995
75. Lou S, Hu Z, Zhang Y, Feng Y, Zhou M, Lv C (2024) Human-cyber-physical system for industry 5.0: a review from a human-centric perspective. IEEE Trans Autom Sci Eng 22:494–511
76. Lu N, Liu L, Yu D, Fu B (2021) Navigating trade-offs in the social-ecological systems. Curr Opin Environ Sustain 48:77–84. https://doi.org/10.1016/j.cosust.2020.10.014
77. Mahajan S, Hausladen CI, Sánchez-Vaquerizo JA, Korecki M, Helbing D (2022) Participatory resilience: surviving, recovering and improving together. Sustain Cities Soc 83, Article ID 103942

78. Maiden N, Lockerbie J, Zachos K, Wolf A, Brown A (2023) Designing new digital tools to augment human creative thinking at work: an application in elite sports coaching. Expert Syst 40(3):e13194. https://doi.org/10.1111/exsy.13194
79. Maitlis S (2005) The social processes of organizational sensemaking. Acad Manag J 48(1):21–49
80. Maitlis S, Christianson M (2014) Sensemaking in organizations: taking stock and moving forward. Acad Manag Ann 8(1):57–125. https://doi.org/10.1080/19416520.2014.873177
81. Malone TW, Malone TW, Crowston K, Crowston K (1994) The interdisciplinary study of coordination. ACM Comput Surv 26(1):87–119. https://doi.org/10.1145/174666.174668
82. Matthias A (2004) The responsibility gap: ascribing responsibility for the actions of learning automata. Ethics Inf Technol 6(3):175–183. https://doi.org/10.1007/s10676-004-3422-1
83. Muhren WJ, Van de Walle B (2010) A call for sensemaking support systems in crisis management. In: Babuška R, Groen FCA (eds) Interactive collaborative information systems. Springer, Berlin, pp 425–452. https://doi.org/10.1007/978-3-642-11688-9_16
84. Muller J (2018) The tyranny of metrics. Princeton University Press, Princeton
85. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, 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(1):143. https://doi.org/10.1186/s12874-018-0611-x
86. Nabizadeh Rafsanjani H, Nabizadeh AH (2023) Towards human-centered artificial intelligence (AI) in architecture, engineering, and construction (AEC) industry. Comput Human Behav Rep 11. https://doi.org/10.1016/j.chbr.2023.100319
87. Neale DC, Carroll JM, Rosson MB (2004) Evaluating computer-supported cooperative work: models and frameworks. In: Proceedings of the 2004 ACM conference on computer supported cooperative work, Chicago, Illinois, USA. https://doi-org.tudelft.idm.oclc.org/10.1145/1031607.1031626
88. Nespeca V, Comes T, Brazier F (2021) A methodology to develop agent-based models for policy design in socio-technical systems based on qualitative inquiry. In: Czupryna M, Kamiński B (eds) Social simulation conference 2021. Springer, Berlin
89. OECD (2019) Recommendation of the council on artificial intelligence. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449
90. Ozmen Garibay O, Winslow B, Andolina S, Antona M, Bodenschatz A, Coursaris C, Falco G, Fiore SM, Garibay I, Grieman K, Havens JC, Jirotka M, Kacorri H, Karwowski W, Kider J, Konstan J, Koon S, Lopez-Gonzalez M, Maifeld-Carucci I, et al (2023) Six human-centered artificial intelligence grand challenges. Int J Hum-Comput Interact 39(3):391–437. https://doi.org/10.1080/10447318.2022.2153320
91. Parasuraman R, Sheridan TB, Wickens CD (2000) A model for types and levels of human interaction with automation. IEEE Trans Syst Man Cybern, Part A, Syst Hum 30(3):286–297. https://doi.org/10.1109/3468.844354
92. Passlack N, Hammerschmidt T, Posegga O (2024) It was(n't) me: vignette experiments on managers' responsibility attribution in AI-advised decision-making. Behav Inf Technol 1(30). https://doi.org/10.1080/0144929X.2024.2431050
93. Paulus D, Fathi R, Fiedrich F, de Walle BV, Comes T (2022) On the interplay of data and cognitive bias in crisis information management. Information systems frontiers. https://doi.org/10.1007/s10796-022-10241-0
94. Ploeg J, Semsar-Kazerooni E, Lijster G, Van de Wouw N, Nijmeijer H (2014) Graceful degradation of cooperative adaptive cruise control. IEEE Trans Intell Transp Syst 16(1):488–497
95. Ponti M, Seredko A (2022) Human-machine-learning integration and task allocation in citizen science. Humanit Soc Sci Commun 9(1):48. https://doi.org/10.1057/s41599-022-01049-z
96. Rahwan I, Cebrian M, Obradovich N, Bongard J, Bonnefon J-F, Breazeal C, Crandall JW, Christakis NA, Couzin ID, Jackson MO (2019) Machine behaviour. Nature 568(7753):477–486
97. Riedl C, De Cremer D (2025) AI for collective intelligence. Collect Intell 4(2), Article ID 26339137251328909
98. Riedl MO (2019) Human-centered artificial intelligence and machine learning. Hum Behav Emerg Technol 1(1):33–36
99. Rittel HWJ, Webber MM (1973) Dilemmas in a general theory of planning. Policy Sci 4(2):155–169
100. Sadek M, Kallina E, Bohné T, Mougenot C, Calvo RA, Cave S (2024) Challenges of responsible AI in practice: scoping review and recommended actions. AI Soc. https://doi.org/10.1007/s00146-024-01880-9
101. Sætre AS, Van de Ven A (2021) Generating theory by abduction. Acad Manag Rev 46(4):684–701
102. Seidel S, Chandra Kruse L, Székely N, Gau M, Stieger D (2018) Design principles for sensemaking support systems in environmental sustainability transformations. Eur J Inf Syst 27(2):221–247. https://doi.org/10.1057/s41303-017-0039-0
103. Sellen A, Horvitz E (2024) The rise of the AI co-pilot: lessons for design from aviation and beyond. Commun ACM 67(7):18–23. https://doi.org/10.1145/3637865
104. Sharma N, Liao QV, Xiao Z (2024) Generative echo chamber? Effect of LLM-powered search systems on diverse information seeking. In: Proceedings of the 2024 CHI conference on human factors in computing systems, Honolulu, HI, USA. https://doi-org.tudelft.idm.oclc.org/10.1145/3613904.3642459
105. Sharoda PA, Reddy MC (2010) Understanding together: sensemaking in collaborative information seeking CSCW
106. Sheridan TB, Verplank WL, Brooks TL (1978). Human/computer control of undersea teleoperators
107. Shneiderman B (2020) Human-centered artificial intelligence: reliable, safe & trustworthy. Int J Hum-Comput Interact 36(6):495–504
108. Shulner-Tal A, Kuflik T, Kliger D (2023) Enhancing fairness perception–towards human-centred AI and personalized explanations understanding the factors influencing laypeople's fairness perceptions of algorithmic decisions. Int J Hum-Comput Interact 39(7):1455–1482. https://doi.org/10.1080/10447318.2022.2095705
109. Sigfrids A, Leikas J, Salo-Pöntinen H, Koskimies E (2023) Human-centricity in AI governance: a systemic approach [Perspective]. Front Artif Intell 6. https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2023.976887
110. Søgaard Jørgensen P, Jansen REV, Avila Ortega DI, Wang-Erlandsson L, Donges JF, Österblom H, Olsson P, Nyström M, Lade SJ, Hahn T (2024) Evolution of the polycrisis: Anthropocene traps that challenge global sustainability. Philos Trans R Soc Lond, B 379(1893), Article ID 20220261

111. Sprenkamp K, Dolata M, Schwabe G, Zavolokina L (2025) Data-driven intelligence in crisis: the case of Ukrainian refugee management. Gov Inf Q 42(1), Article ID 101978. https://doi.org/10.1016/j.giq.2024.101978
112. Stapleton L, Lee MH, Qing D, Wright M, Chouldechova A, Holstein K, Wu ZS, Zhu H (2022) Imagining new futures beyond predictive systems in child welfare: a qualitative study with impacted stakeholders. In: Proceedings of the 2022 ACM conference on fairness, accountability, and transparency, Seoul, Republic of Korea. https://doi-org.tudelft.idm.oclc.org/10.1145/3531146.3533177
113. Stasser G, Titus W (1985) Pooling of unshared information in group decision making: biased information sampling during discussion. J Pers Soc Psychol 48(6):1467–1478. https://doi.org/10.1037/0022-3514.48.6.1467
114. Stewart DD, Stasser G (1995) Expert role assignment and information sampling during collective recall and decision making. J Pers Soc Psychol 69(4):619–628
115. Stone P, Kaminka G, Kraus S, Rosenschein J (2010). Ad hoc autonomous agent teams: collaboration without pre-coordination
116. Suchan J, Bhatt M, Varadarajan S (2021) Commonsense visual sensemaking for autonomous driving – on generalised neurosymbolic online abduction integrating vision and semantics. Artif Intell 299. https://doi.org/10.1016/j.artint.2021.103522
117. Sun J (2022) Investigating explainability of generative AI for code through scenario-based design. 27th international conference on intelligent user interfaces. https://doi.org/10.1145/3490099.3511119
118. Tausch A, Kluge A (2022) The best task allocation process is to decide on one's own: effects of the allocation agent in human–robot interaction on perceived work characteristics and satisfaction. Cogn Technol Work 24(1):39–55. https://doi.org/10.1007/s10111-020-00656-7
119. Tetlock PE (2003) Thinking the unthinkable: sacred values and taboo cognitions. Trends Cogn Sci 7(7):320–324. https://doi.org/10.1016/S1364-6613(03)00135-9
120. Thompson J (2021) Mental models and interpretability in AI fairness tools and code environments. In: Stephanidis C, Kurosu M, Chen JYC, Fragomeni G, Streitz N, Konomi S, Degen H, Ntoa S (eds) HCI international 2021 - late breaking papers: multimodality, eXtended reality, and artificial intelligence. Springer, Cham
121. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, Moher D, Peters MDJ, Horsley T, Weeks L (2018) PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 169(7):467–473
122. UNDRR (2025) Special report on the use of technology for disaster risk reduction. https://www.undrr.org/quick/94683
123. UNESCO (2022) Recommendation on the ethics of artificial intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000381137
124. Vallor S (2015) Moral deskilling and upskilling in a new machine age: reflections on the ambiguous future of character. Philos Technol 28(1):107–124. https://doi.org/10.1007/s13347-014-0156-9
125. Van Berkel N, Bellio M, Skov MB, Blandford A (2023) Measurements, algorithms, and presentations of reality: framing interactions with AI-enabled decision support. ACM Trans Comput-Hum Interact 30(2):1–33
126. van Berkel N, Tag B, Goncalves J, Hosio S (2022) Human-centred artificial intelligence: a contextual morality perspective. Behav Inf Technol 41(3):502–518. https://doi.org/10.1080/0144929X.2020.1818828
127. van Leersum CM, Maathuis C (2025) Human centred explainable AI decision-making in healthcare. J Responsib Technol 21, Article ID 100108. https://doi.org/10.1016/j.jrt.2025.100108
128. Verma H, Mlynar J, Schaer R, Reichenbach J, Jreige M, Prior J, Evéquoz F, Depeursinge A (2023) Rethinking the role of AI with physicians in oncology: revealing perspectives from clinical and research workflows. CHI Human Factors
129. Volkema RJ (1983) Problem formulation in planning and design. Manag Sci 29(6):639–652. https://doi.org/10.1287/mnsc.29.6.639
130. Webster M (2008) Incorporating path dependency into decision-analytic methods: an application to global climate-change policy. Decis Anal 5(2):60–75
131. Weick K, Sutcliffe K (2007) Managing the unexpected: resilient performance in an age of uncertainty, second edition. http://dl.acm.org/citation.cfm?id=1408051
132. Weick KE (1993) The collapse of sensemaking in organizations: the Mann gulch disaster. Adm Sci Q 38(4):628–652
133. Weick KE (1995) Sensemaking in organizations, vol 3. Sage, Thousand Oaks
134. Weick KE (2015) Ambiguity as grasp: the reworking of sense. J Conting Crisis Manag 23(2):117–123
135. Wiener N (1948) Cybern Sci Am 179(5):14–19
136. Woolley AW, Chabris CF, Pentland A, Hashmi N, Malone TW (2010) Evidence for a collective intelligence factor in the performance of human groups. Science 330(6004):686–688. https://doi.org/10.1126/science.1193147
137. Yazdanpanah V, Gerding EH, Stein S, Dastani M, Jonker CM, Norman TJ (2021) Responsibility research for trustworthy autonomous systems AAMAS. Online
138. Zhang R, Flathmann C, Musick G, Schelble B, McNeese NJ, Knijnenburg B, Duan W (2024) I know this looks bad, but I can explain: understanding when AI should explain actions in human-AI teams. ACM Trans Interact Intell Syst 14(1):6. https://doi.org/10.1145/3635474
139. Zhou J, Zhou Y, Wang B, Zang J (2019) Human–Cyber–Physical Systems (HCPSs) in the context of new-generation intelligent manufacturing. Engineering 5(4):624–636. https://doi.org/10.1016/j.eng.2019.07.015
140. Zwitter A (2024) Cybernetic governance: implications of technology convergence on governance convergence. Ethics Inf Technol 26(2):24


## Publisher's note