



Delft University of Technology

Document Version

Final published version

Licence

Dutch Copyright Act (Article 25fa)

Citation (APA)

Gadiraju, U., & Balayn, A. (2025). The Enterprising and Elusive Prospects of Human-AI Collaboration. In S. Sadiq (Ed.), *Enterprise AI* (pp. 211-243). Springer. https://doi.org/10.1007/978-3-032-01940-0_7

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

In case the licence states "Dutch Copyright Act (Article 25fa)", this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

Chapter 7

The Enterprising and Elusive Prospects of Human-AI Collaboration



Ujwal Gadiraju and Agathe Balayn

Abstract The exponential advances in generative AI and agentic technologies have presented organizations with an unprecedented opportunity to seek and attain competitive advantages. Some aim to do so through product and service innovation, while others have begun to pursue operational efficiency and productivity. There is a pivotal role that human-AI collaboration can play in enterprise artificial intelligence design, development, and deployment today. In this chapter, we synthesize developments in human-AI collaboration over the last decade and explore how organizations can design AI systems that augment rather than replace human capabilities to achieve optimal experiential and performance-related outcomes for different stakeholders. Drawing from empirical work across different domains, we analyze the key factors that shape trust and reliance and determine successful human-AI collaboration. This includes various human factors, task factors, and AI system factors and the complex interplay between them in different configurations. This synthesis reflects the importance of broadening the spectrum of metrics for evaluating human-AI collaboration (e.g., by considering stakeholder values). We discuss promising ideas to address common challenges such as under-reliance or over-reliance on AI systems. We argue for broadening the lens of human-AI collaboration to consider the AI supply chain and the underlying value chains to ensure the responsible design of enterprise AI. Our insights collectively suggest that enterprise AI will benefit from the human-centered approach while creating collaborative workflows and human-AI configurations that can augment and complement humans with the computational power and scalability of AI.

U. Gadiraju (✉)
Delft University of Technology, Delft, The Netherlands
e-mail: u.k.gadiraju@tudelft.nl

A. Balayn
Microsoft Research, New York, NY, USA
e-mail: balaynagathe@microsoft.com

7.1 Introduction

The fundamental essence of *human-AI collaboration* is not novel. Similar ideas were proposed and discussed as early as in 1960 by Licklider [1]. Over the last few years, we have witnessed increasing adoption of AI systems across different domains and around the world. Opportunities created by the potential of AI technologies have been met with euphoric enthusiasm and scathing skepticism in equal parts—and for good reason. On the one hand, advances in machine learning and AI have led to remarkable breakthroughs. In healthcare, systems like AlphaFold have revolutionized protein structure prediction, leading to an unprecedented acceleration in drug discovery and development [2, 3]. There is growing evidence suggesting the potential of AI systems for question-answering related to medicine, medical diagnosis, and precision medicine [4–7]. AI systems have also been built to good effect in pursuit of climate modeling [8] and conservation of biodiversity [9]. AI-powered chatbots have transformed customer service [10] and are being adopted in healthcare and education to solve a variety of challenges [11–14]. From transportation to finance and entertainment, the list of positive footprints of AI technologies goes on. On the other hand, the frantic adoption of AI systems across the world has also led to unforeseen and unfortunate consequences. These have been well-documented by media, journalists, and other efforts that have led to systematically capturing such “*AI incidents*” in a database to prevent repeated AI harms or failures [15]. In this context, researchers and practitioners across different communities have proposed frameworks, tools, and methods to build responsible, ethical, trustworthy, safe, robust, meaningful, transparent, and *human-centered* AI systems—with often overlapping goals and objectives [16–21].

Many Flavors of Human-AI Collaboration and Teaming

The potential of AI systems to support, complement, and augment the skills and capabilities of humans across different tasks has led to widespread applications in human-AI collaborative configurations with varying effectiveness [22–27]. For instance, Sowa et al. explored synergies between human workers and AI in managerial tasks, identifying an increased productivity due to human–AI collaboration [28]. There is a growing footprint of AutoAI or AutoML in data science, wherein data scientists are aided by automated workflows that simplify their practices [29]. Jiang et al. explored task delegability in human-AI collaboration in the context of qualitative analysis and found that the design of AI assistance that honor serendipity and human agency are necessary for meaningful collaboration [30]. More recently, Rastogi et al. highlighted the complementary strengths of humans and generative models in collaborative auditing of large language models (LLMs) [31].

Prior work has explored the prospect of integrating AI systems into human teams to achieve shared goals [32–35]. In contrast to human-AI collaboration where humans and AI work together on a given task but maintain relatively distinct roles and responsibilities, human-AI teaming entails a dynamic relationship, resembling

the interaction between human team members with AI as an active participant.¹ While human-AI collaboration has largely been associated with the notion of using AI as a functional assistant, human-AI teaming predominantly corresponds to integrating AI as an adaptive and autonomous partner in a shared process. These distinctions, however, have not always been made consistently in existing literature spanning both conceptualizations. Thus, there is some overlap across these realms of work. In the context of multiplayer online games, Zhang et al. found that people reported mixed feelings toward AI teammates but exhibited a generally positive attitude toward future collaboration with AI teammates [38]. In other empirical work, Zhang et al. assessed the effects of teammate identity (comparing human versus AI) and teammate performance (comparing low-performing versus high-performing AI) on human-AI cooperation [39].

He et al. explored the potential of using LLM agents with humans-in-the-loop for planning tasks and found that LLM agents can work effectively when a high-quality plan and necessary user involvement in execution are available and that users can easily mistrust LLM agents with plans that seem plausible [40]. In the evolving area of explainable AI, recent evidence has shown that the use of LLM agents can result in amplifying over-reliance on AI systems by creating an illusory depth in explanations [41]. Over the last decade, a number of empirical qualitative and quantitative studies have contributed to a wealth of knowledge around how we can design for better human-AI collaboration and build effective human-AI teams [42, 43].

Considering that organizations and companies are invested in optimizing their business and operations with AI [44], it is important to address and understand the potential of human-AI collaboration for enterprise AI. Recent work at the intersection of HCI and AI has identified *calibrated trust* as an important need for appropriate interaction with assistive AI [45]. Some pivotal questions that arise in this context are:

- How can AI systems assist and support human decision-making while optimizing for trade-offs that affect enterprises?
- How can we foster appropriate trust and reliance on collaborative decision-support systems in enterprises?

This chapter unpacks recent findings in human-AI collaboration and reflects on future directions for enterprise AI. To this end, we first present an overview of the notions of trust and reliance on AI systems that have been proposed in existing literature and then synthesize findings around factors that shape human-AI

¹ Note that there are further conceptualizations of combining human and machine intelligence, such as *hybrid intelligence* [36, 37]. Though relevant, this remains beyond the scope of our chapter.

collaboration. Note that our view on *trust* considers it as a mechanism that can foster effective collaboration between humans and AI systems, as opposed to considering *trust* as an end goal that we should strive to optimize for [46].

7.2 Trust and Reliance on AI Systems

In their impactful work on *trust in automation*, Lee and See defined trust as the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability [47]. They argued that subjective trust can guide and influence objective behavior and interactions (i.e., reliance) with an automated system. Hoff and Bashir reviewed empirical research on factors that influence trust in automation, published between 2002 and 2013, and proposed a three-layered interdependent trust model [48]. This consists of (i) *dispositional trust* that represents an individual's overall tendency to trust automation, independent of context or a specific system; (ii) *situational trust* that represents the variability in trust depending on different contexts; and (iii) *learned trust* that represents an individual's assessment of a system based on past experience or the current interaction.

More recently, the notion of regulating one's *trust in an AI system* and thereby informing their reliance on said system, preventing disuse and over-use (i.e., under-reliance and over-reliance), and relying on AI advice only when it is credible and warranted has been referred to as the pursuit of facilitating *appropriate trust*. Evidence from different empirical studies suggests that all three layers of the interdependent trust model [48] have been shown to influence appropriate trust. Over the last decade, researchers in HCI and AI have explored a number of methods to foster appropriate trust and reliance in human-AI interactions to pave the way for optimal human-AI collaboration.

Mehrotra et al. carried out a comprehensive and systematic review of the growing literature on fostering appropriate trust in AI [49]. In their work, the authors also shed light on different concepts related to appropriate trust, including warranted trust [50], justified trust [51], contractual trust [52, 53], and calibrated trust [54, 55]. Jacovi et al. proposed a formalization of trust and how to design trustworthy AI [52]. The authors discuss the possible causes of warranted trust as intrinsic reasoning and extrinsic behavior. More recently, Schlicker et al. proposed a two-level trustworthiness assessment model (TrAM) that applies psychological models to evaluate system trustworthiness [56] and distinguishes system characteristics, actual and perceived trustworthiness, individual standards, and cues. Miller proposed methods to measure perceived and demonstrated trust [57]. Several empirical works studying human-AI interaction and collaboration have operationalized and leveraged these underpinnings of trust in AI to assess novel methods and their effectiveness in fostering appropriate trust and reliance on AI systems [25, 58]. In particular, there is a growing body of work exploring the potential of AI to aid human decision-

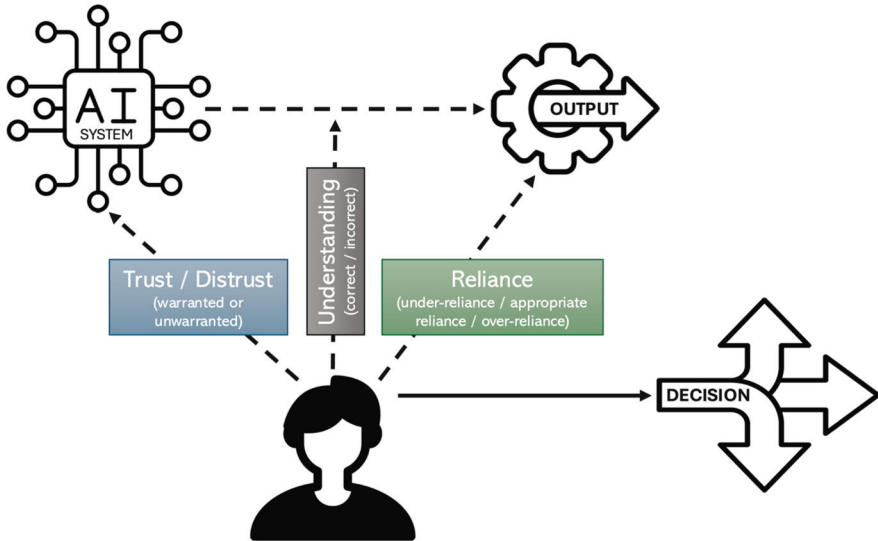


Fig. 7.1 Relation between *trust*, *reliance*, and the appropriateness of their reliance in human-AI collaboration contexts (e.g., decision-making). Users may trust or distrust an AI system, and this may either be warranted or unwarranted. Their understanding of the process through which the AI system reaches its output, or the AI system’s rationale behind its advice, can play an important role in shaping their reliance on the AI output (e.g., using AI assistance in a decision-making task)

making [59–63]. In fact, optimizing decision workflows has been identified as an avenue that bears immense potential for enterprise AI [44].

In summary (see Fig. 7.1), while trust plays a central role in shaping the effectiveness of decision workflows, it is ill advised to pursue trust as a goal in and of itself. It is far more beneficial to view fostering *appropriate* trust as a means to facilitate effective human-AI collaboration and team performance [41, 46, 64]. In that vein, it is worth asking whether appropriate trust is sufficient to achieve effective human-AI collaboration. It is now understood that although (appropriate) trust plays a foundational role in human-AI collaboration, it does not solely lead to effective human-AI collaboration (i.e., it is necessary but not sufficient). Several critical factors influence the effectiveness of human-AI collaboration—complementary capabilities, feedback mechanisms, and technical qualities of the AI system, among others. For example, consider the prospect of complementary capabilities [65]. In industrial or manufacturing settings [66], predictive maintenance systems proficiently identify patterns in different types of data that predict equipment failure, while human operators contribute contextual knowledge about specific machines and their operating conditions. In such a collaborative context where appropriate trust on an AI system exists but there is a lack of complementary capability, benefits of the collaboration cannot be reaped.

7.3 Key Factors That Shape Outcomes in Human-AI Collaboration

In this section, we present an overview of the multitude of factors that have been shown to influence user trust, as well as experiential outcomes and team performance in human-AI collaboration. These factors (presented in Table 7.1), together with human-AI configurations (see Table 7.2), blend together in a complex interplay that shapes team performance and the perceived experiences.

7.3.1 Human Factors

A number of human factors and individual traits have been shown to influence and shape both experiential outcomes and performance outcomes in human-AI interaction and collaboration. These factors need to be considered while designing decision workflows or interactions with AI systems in the context of enterprise AI.

AI literacy, i.e., the proficiency of humans in different domains of AI that can enable the efficient, meaningful, and ethical use of AI technologies, has been widely identified as being a crucial factor in regulating AI adoption and

Table 7.1 An overview of the factors that have been shown to shape trust and affect human-AI collaboration in prior work. References serve as examples of works that have studied corresponding factors. Note that this is not an exhaustive summary of all relevant factors surfaced in existing literature but a representation of some key factors that have been methodically studied and are well understood to play an important role

Human factors	Task factors	AI system factors
AI literacy [67–69]	Complexity [70]	Interpretability and
Domain knowledge [71]	Uncertainty [63]	explainability [72]
Domain expertise [73]	Stakes [61, 74]	Accuracy [75, 76]
Mental models [77]	Task context (e.g., time	Reliability [78, 79]
Prior experience [80, 81]	constraints, pressure [82, 83])	Trustworthiness [57]
Cognitive biases [80, 84]		Interface (e.g., modality,
User expectations [85, 86]		feedback, accessibility) [87, 88]
Culture and values [89, 90]		

Table 7.2 Different aspects of human-AI configurations that have been shown to shape experiential and performance-related outcomes in human-AI collaboration

Human-AI configurations
Distribution of control [18]
Individual versus collaborative use [116, 129, 130]
Organizational dynamics, policies [131]
Interdependency [132]

fostering appropriate trust and reliance on AI systems, consequently paving a way for effective human-AI collaboration. Cetindamar et al. identified four sets of capabilities associated with AI literacy—(i) technology-related, (ii) work-related, (iii) human-machine-related, and (iv) learning-related capabilities—and pointed out the importance of operationalizing AI literacy for non-AI professionals [91]. Over the last few years, there have been several efforts to develop AI literacy scales [68, 69, 92, 93] to understand the role and assess the impact of AI literacy on human-AI collaboration on one hand and efforts to increase AI literacy in society to help people effectively interact with and critically evaluate AI systems on the other [67, 94–96].

Hoffman et al. presented a broad overview of *mental model* elicitation methods and their application in the context of explainable AI [97]. The authors highlighted that user mental models represent their understanding of an AI system and its context, including the task, goals, and current situation. Bansal et al. found that human *mental models of AI capabilities* and their awareness of when to accept or override an AI system's recommendation (based on their mental models) affect the resulting team performance [77]. Andrews et al. proposed a conceptual model for the application of literature on “shared mental models” in human teams to the context of human-AI collaboration [98]. Gero et al. studied people's mental models of AI agents in a cooperative word-guessing game [99]. The authors found that people have existing intuitions about how AI systems work and that such intuitions can be disrupted when they encounter specific AI systems. They also characterized the capability of people to revise their mental models of the AI system in the face of encountering anomalies.

Researchers have also studied how *user expectations* and various AI metaphors influence and shape experiential and performance outcomes in human-AI collaboration. Khadpe et al. found that conceptual metaphors used to describe AI systems can shape user expectations and behavior in human-AI collaboration [85]. Similarly, in the context of conversational crowdsourcing, Jung et al. found that different human and non-human metaphors can affect worker engagement, their perceived cognitive load in various tasks, their intrinsic motivation, and their trust in the agents [86]. Cabrera et al. found that showing users behavior descriptions of AI systems, i.e., details of how AI systems perform on subgroups of instances, can improve the performance of human-AI teams across different tasks [100]. Furthermore, Nourani et al. uncovered and synthesized the role of various *experiential biases* in shaping user mental models and outcomes in their interactions or collaboration with AI systems [101]. For instance, first impressions of AI systems have been shown to weigh heavily on influencing user expectations and reliance behavior on AI systems [102]. However, some prospects for the evolution of perceptions beyond first impressions have also been found [81] in different contexts.

Domain knowledge, experience, and expertise have all been identified as factors that affect human-AI collaboration. Zhang et al. explored how users trust and rely on an AI assistant that performs with different levels of expertise relative to themselves, ranging from completely overlapping expertise to perfectly complementary expertise. The authors found that participants were able to perceive when

the assistant was an expert or non-expert within the same task and calibrate their reliance on the AI to improve team performance [73]. To optimize human-AI team performance, Wilder et al. carried out a systematic exploration of how machine learning systems can be trained to complement human reasoning [103]. Through a controlled study in a human-AI collaboration task, Nourani et al. explored the role of domain knowledge in establishing trust and susceptibility to the influence of first impressions on user trust [71]. The authors found a differing impact on domain experts versus novice users. Encountering errors early on caused negative first impressions for some domain experts, negatively impacting their trust over the course of interactions. In contrast, novice users were found to suffer from over-reliance due to their lack of proper knowledge to detect errors. Inkpen et al. explored the interplay between user expertise and complementarity of AI systems on human-AI collaboration [60]. The authors found that novice users demonstrated improvements in performance by relying on AI advice, highly proficient users were able to discern when to follow AI advice and typically maintained or improved their performance as a result of the collaboration, and that mid-performers (with a similar level of accuracy to the AI) were most variable in terms of whether the AI advice helped or hurt their performance. However, there is growing evidence suggesting that achieving optimal human-AI team performance can be challenging. Agarwal et al. found that providing expert radiologists with AI predictions does not always improve performance, whereas contextual information does. Radiologists do not realize the gains from AI assistance because of errors in belief updating—they underweight AI predictions and treat their own information and AI predictions as statistically independent. Unless these mistakes can be corrected, the optimal human-AI collaboration design delegates cases either to humans or to AI, but rarely to AI-assisted humans.

That humans are susceptible to a plethora of *cognitive and behavioral biases* that influence how different tasks are carried out is a well-established fact [104, 105]. Prior work has unveiled how biases can surface in different contexts of human-AI collaboration. Echterhoff et al. explored a specific type of anchoring bias in which decision-makers are anchored by their own recent decisions and found that it is possible to combine the strengths of AI systems and humans to enable fairer decisions in sequential decision-making tasks [106]. Through a controlled study, Nourani et al. found that users who observed strengths of an AI system early on in their interactions were more prone to automation bias and made significantly more errors due to positive first impressions of the system and that users who encountered weaknesses of an AI system earlier made significantly fewer errors since they tended to rely more on themselves [80]. Building on work from the field of cognitive science, Rastogi et al. proposed a general framework to account for cognitive biases in the human-AI collaborative decision-making setting and mitigate their negative effects on team performance [107]. Based on their empirical exploration of human-AI collaboration, Jacobsen et al. discussed the gap between perceived and measured task effectiveness [108]. He et al. showed that the meta-cognitive bias of the Dunning-Kruger Effect can have a detrimental effect on team performance in human-AI collaboration [84]. The authors found that

participants who overestimated their performance tended to exhibit under-reliance on AI systems, which in turn hindered optimal team performance.

Erlei et al. found that humans tend to overwrite their economic self-interest to avoid bargaining with AI systems [64], and also that humans consistently violate the choice independence assumption when delegating predictions to a superior AI system [213]. Vodrahalli et al. highlight the importance of modeling human interaction when designing AI systems for human use [24]. Drawing from empirical findings across different tasks, they reflected on the importance of jointly optimizing the human-AI collaborative system instead of the predominant paradigm of optimizing the AI model alone. Biswas et al. found that bilingual writers demonstrated biased behavior when interacting with an AI co-writing agent, wherein prior exposure to a Spanish LLM reduced subsequent utilization of a better-performing English LLM in a persuasive charitable giving task [109].

There is an increasing understanding and recognition of the role of *culture* and human *values* in shaping human interactions with AI systems [89, 110–113]. Considering the progress in AI regulation around the world and an echoing rally to ensure AI systems adhere to certain values, van de Poel identified the complex challenges and opportunities of embedding values in AI systems [89]. In line with the grand objective of aligning AI with human values [114], Hendrycks et al. proposed a method to assess a language model’s knowledge of basic concepts of morality [112]. Seth et al. found a significant variation across different language models in their ability to infer artifacts across regional sub-cultures in India [113].

7.3.2 Task Factors

Different task characteristics have been shown to influence both experiential and performance-related outcomes in human-AI collaboration [115, 116]. This section presents an overview of such task factors.

Lai et al. surveyed prior work in human-AI collaboration and identified four dimensions of task characteristics that varied in the tasks: (1) *task risk* or *stakes* (e.g., high, low, or artificial stakes), (2) *required expertise* in the task, (3) *decision subjectivity*, and (4) AI for *emulation* versus *discovery* [61]. A recent empirical study by Baron et al. explored the impact of stakes of the decision (high versus low) and the decision source (human versus AI) on user perceptions of explanations for AI systems in medical contexts [74]. The authors found that the effectiveness of explanations may not depend on the type of explanation but instead on the stakes and decision source.

Salimzadeh et al. argued that the variety in decision-making tasks empirically studied in the context of human-AI collaboration and their operationalization has led to an opacity around how findings from one task or domain carry forward to another [70]. The authors proposed measuring *task complexity* to tackle the problem of under-specification and facilitate comparison across empirical research in human-AI collaboration. Tomsett et al. proposed that the objectives of imbuing

AI systems with interpretability and uncertainty-awareness can allow for rapid trust calibration among decision-makers, by facilitating the understanding of an AI system's limitations and likely failures [117]. Prabhudesai et al. explored how and why lay decision-makers utilize ML *uncertainty* information in their decision process [118]. Through a qualitative study, they found that communicating uncertainty about a machine learning model's predictions forced participants to slow down and think analytically about their decisions. Other empirical work by Salimzadeh et al. found that when users are presented with complex and uncertain tasks, they tended to rely more on AI systems while demonstrating lower levels of appropriate reliance compared to tasks that were less complex and uncertain [63].

Cao et al. empirically studied human decision-making behavior and task performance under *time pressure* that is either induced from limited initial observation time or final decision time (i.e., the time to weigh an AI's suggestion before reaching a collective human-AI team answer) [83]. The authors found that users tended to follow AI suggestions when they were provided with a longer final decision time. In a similar context of studying human-AI collaboration under time pressure, Swaroop et al. found that different types of AI assistance have different accuracy-time trade-offs when people are under time pressure compared to no time pressure [82].

7.3.3 AI System Factors

This section presents an overview of how different attributes of an AI system,² ranging from its interpretability to accuracy and reliability, play a role in shaping performance and experiences in human-AI collaboration.

Several efforts have been made to communicate AI system predictions, advice, or output to human stakeholders—from stating system accuracy to generating *explanations*—with varying outcomes [75, 76, 119]. Yin et al. found that users' trust in a model is affected by both its *stated accuracy* and its observed accuracy and that the effect of stated accuracy can change depending on the observed accuracy [75]. However, He et al. explored the impact of communicating system accuracy on human-AI collaboration outcomes and found that explaining the stated accuracy of a system using analogies failed to help users rely on the AI system appropriately [76]. Rechkemmer et al. investigated user trust in AI when multiple performance indicators are presented at the same time [120]. The authors found that while the level of *model confidence* significantly affects people's belief in model accuracy, both the model's stated and observed accuracy generally have a larger impact on people's willingness to follow the model's predictions.

² In this chapter, we use the phrase “an AI system” interchangeably with an AI-based system, a machine learning model, a decision-support system, an AI assistant, or an algorithmic decision aid.

Senoner et al. carried out experiments with real-world visual inspection tasks in the domain of manufacturing and medicine to understand the impact of explainable AI aids on human-AI collaboration [72]. They found that domain experts performing the tasks demonstrated a significant improvement in their performance when they were supported by explainable AI aids in the form of visual heatmaps. Ahn et al. investigated how interpretability and *outcome feedback* influence users' task performance and trust in AI [121]. In contrast to other findings suggesting that interpretability can drive trust, the authors found that interpretability led to no robust improvements in trust and that outcome feedback had a significantly greater and more reliable effect. Similarly, in a controlled experiment with AI practitioners, Kaur et al. found that participants spent significantly less time on tasks and were far less accurate about the data and model when they were given access to interpretability tools [122]. This complements earlier results showcasing how data scientists tend to over-trust and misuse interpretability tools [123].

Through mixed-method user studies, Bansal et al. found complementary improvements resulting from human-AI collaboration and also that such benefits were not increased by using explanations. They found that explanations instead increased the likelihood of humans accepting the AI system's recommendation, regardless of its correctness [59]. In other work, Bansal et al. explored the influence of *technical updates* to an AI system in a human-AI collaboration context. They found that although updates can be made to increase the system's performance, they can also lead to behavioral changes that deviate from users' prior experiences and confidence in the AI system's inferences, thereby having a negative impact on team performance [124].

Early work by Fan et al. showed that the knowledge of *AI reliability* among users has a significant impact on shaping user trust [78]. Appelganc et al. explored the reliability that participants required to perceive support agents (either human or AI systems) as being "highly reliable" and how this differed across contexts of working together or being evaluated by the agent [79]. Through an empirical study, the authors found that the required reliability in the context of working together did not differ across the humans or AI. However, participants required that AI be more reliable than a human when used to evaluate themselves.

The *modality of interaction* with AI systems have also been shown to influence performance outcomes and interaction experiences in human-AI collaboration [88, 125–127]. Le et al. compared spoken and written interactions with a natural language dialogue system using two real information-retrieval systems. The authors found that the spoken interaction mode was less efficient but promoted collaboration [127]. Zhang et al. highlighted the trade-offs between AI's verbal communication and visual communication based on their findings from an empirical study exploring human-AI team compositions [125]. Through an empirical study, Liu et al. found that the visual-auditory bimodality was superior to single visual modality and single auditory modality for older adults based on the perceptions of social presence, attitudinal outcomes, and acceptance [126]. The authors observed that older adults perceived greater social presence, perceived enjoyment, satisfaction, and acceptance with text feedback than without in their interactions with

AI assistants. He et al. explored the impact of a conversational XAI interface on users' understanding of the AI system, their trust, and reliance on the AI system in comparison to an XAI dashboard [128]. The authors found that the conversational XAI interface could bring about a better understanding of the AI system among users and lead to higher user trust. However, users of both the XAI dashboard and conversational XAI interfaces demonstrated over-reliance on the AI system, and enhancing the conversations by using agents amplified this over-reliance.

7.3.4 Human-AI Configurations

This section synthesizes findings from different human-AI configurations and workflows for human-AI collaboration, including those within organizations, given the widespread adoption of enterprise AI (as shown in Table 7.2).

Wang et al. synthesized six different configurations in human-AI collaboration based on varying levels of human agency: (i) AI suggests and humans decide (*verification*); (ii) humans provide feedback and AI optimizes (*supervision*); (iii) humans and AI systems work independently for certain tasks, sharing a portion of the decision-making process (*cooperation*); (iv) AI decides and humans control (*intervention*); (v) AI decides and humans can exit (*rejection*); and (vii) AI adapts and human input is eliminated (*regulation*) [132].

In a user study with veterinary radiologists, Fogliato et al. explored the effects of providing AI assistance at the start of a diagnostic session in radiology versus after the radiologists made provisional decisions [133]. The authors found that participants who were asked to register provisional responses in advance of reviewing AI inferences were less likely to agree with AI, regardless of whether the advice was accurate. The authors also found that in instances of disagreement with AI, they were less likely to seek the second opinion of a colleague. In the context of generating personas, Shin et al. explored human-AI workflows that differently delegate subtasks in user research between human experts and LLMs [134]. The authors found that better results (i.e., more representative and empathy evoking) are achieved when leveraging human skill in grouping user data by their key characteristics and exploiting LLMs for summarizing pre-grouped data into personas. Wu et al. proposed the notion of LLM chaining, where the output of one step becomes the input for the next, aggregating potential gains along different steps [135]. In a user study exploring LLM chaining, the authors found that users leveraged sub-tasks to calibrate model expectations, compared and contrasted alternative strategies by observing parallel downstream effects, and debugged unexpected model outputs by testing sub-components of a chain. Allen et al. identified junctures in typical crowdsourcing workflows at which the introduction of LLMs could lead to benefits [136]. The authors proposed means to augment existing design patterns for crowd work. In other work, Wu et al. showed that replicating crowdsourcing pipelines can help us understand the relative LLM strengths on different tasks (via performance comparison on sub-tasks) and the

potential of LLMs in complex tasks, where they can complete part of the tasks alongside humans [137].

Cai et al. explored the context of human-AI collaborative decision-making and carried out a qualitative study with pathologists. The authors interviewed them before, during, and after being presented deep neural network (DNN) predictions for prostate cancer diagnosis to learn the types of information that they desired about the AI assistant [129]. The authors found that clinicians desired upfront information about the known strengths and limitations of AI, its subjective point of view, and its overall design objective. Similarly, Salimzadeh et al. highlighted the importance of considering task characteristics when designing human-AI collaborative systems where peers leverage AI assistance [116].

Through a two-year-long case study of an organization in the international maritime trade that introduced automated algorithmic support for data analysis and prediction work, Gronsund et al. explored how configurations of humans and AI systems evolved as the firm adopted AI [131]. The authors found that the notions of automation and augmentation are intertwined in the evolution of organizations as they adopt AI. They highlighted the strategic importance of a *human-in-the-loop* pattern for organizational reflexivity to ensure that the performance of AI meets the requirements of the organization.

7.4 Toward Effective Human-AI Collaboration for Enterprise AI

Different research communities at the intersection of AI and HCI have tirelessly pursued the conceptualization, operationalization, and development of methods to foster appropriate reliance in human-AI interaction and collaboration [62, 138–142]. Researchers and practitioners have proposed a variety of approaches and empirically evaluated them in different domains [143–145]. For the most part, it can be argued that the pursuit of appropriate reliance is an unsolved challenge and has proved to be an elusive goal. Among other reasons, this is primarily due to conflicting findings across different studies (e.g., the early potential and promise of AI explanations as a general vehicle to foster appropriate reliance and shape positive outcomes has been refuted due to mounting evidence that suggests more nuance [146–154]), the unclear transferability of findings across empirical work in this area, and questionable ecological validity.³ While there is a lot of ground that remains to be covered in our pursuit of facilitating appropriate reliance on AI, there have been some noteworthy efforts that have sown some seeds of promise. We discuss these ideas below.

³ *Ecological validity* refers to the extent to which experimental findings can generalize to the real world beyond the experimental frame [155, 156].

Plausibility, Verifiability, Bounded Rationality, and the Compelling Argument of ‘Evaluative AI’ In their recent work, Fok et al. argued that explanations are only useful to the extent that they allow human decision-makers to verify the correctness of the AI’s prediction [157]. The authors highlight that prior works have found that in many decision-making contexts, AI explanations do not facilitate such verification. Kaur et al. investigated the extent to which AI practitioners apply *bounded rationality* when using interpretability tools and whether the application of bounded rationality helps or hurts.⁴ The authors argue that whether the outcomes of bounded rationality are good or bad is dependent on the heuristics that people apply to select a “good enough” option. When such heuristics are inaccurate, then bounded rationality can lead to inaccurate decisions and propagate harm [122]. Miller proposed a conceptual framework called *evaluative AI* to support explainable decision support while tackling the challenges of over- and under-reliance on AI [158]. The author proposed this as a machine-in-the-loop paradigm in which decision support tools provide evidence for and against decisions made by the users, rather than providing recommendations for users to accept or reject. This is aligned with findings that suggest encouraging critical thinking and consideration among users is fundamentally important to help calibrate trust and foster appropriate reliance on AI [159, 160].

Considering Trust Development Over Time and in Context There has been limited consideration of trust development through interactions with AI that are spread over a period of time [81, 161]. In their recent qualitative work, Kahr et al. addressed the often neglected perspective of trust development over time [162]. The authors investigated a real-world case in which logistics experts had worked with an AI system for many years. They found that despite clearly identifying AI system imperfections, experts developed trust over time. Inconsistencies and frequent efforts to improve the AI system disrupted trust development, hindering control, transparency, and understanding of the system. The authors also found that despite the overall trustworthiness, experts overrode correct AI decisions to protect their colleagues’ well-being.

Fine-Grained Notions of Appropriate Reliance In the context of AI-assisted decision-making in composite fact-checking tasks, He et al. investigated the impact of a novel multi-step decision workflow on user reliance behaviors [163]. Their findings demonstrate that human-AI collaboration with the proposed decision workflow can outperform one-step collaboration in specific contexts (e.g., when advice from an AI system is misleading). Analyzing appropriate reliance at fine-grained levels (i.e., at the sub-step level in decomposed tasks) revealed that such workflows can be effective when users demonstrate a relatively high consideration

⁴ As described in [122], *bounded rationality* is an inherent feature of human decision-making; people can seldom consider the utility of all possible choices before reaching a decision, since that would lead to information overload, decision fatigue, and decision paralysis.

of the intermediate steps. The authors discuss and highlight that there is no one-size-fits-all decision workflow that can help obtain optimal human-AI collaboration.

7.4.1 *Downstream AI Users: Broadening the Lens of “Human-AI Collaboration” to Consider Decision-Subjects*

Other stakeholders can be directly affected as a result of end-users collaborating with a given AI system. Such stakeholders are referred to as *decision-subjects*. For an overview of this, consider the inferences of classic AI systems that are based on data samples fed as input to these systems. The outputs of such systems shape the final decisions made by end-users and can (positively or negatively) affect decision-subjects represented by the input samples. Whether the decision-subject accepts a resulting decision or decides to contest it [164] might not only depend on their satisfaction with the decision but also on their appreciation of the user-AI collaboration, their judgment of the AI system itself, and their trust. In turn, whether a decision-subject appropriately accepts or contests a decision has repercussions not only for themselves but also for the AI deployer who might have to conduct additional work (e.g., to verify the decision, to rebut the decision-subject’s contestation, or pursue some action). Particularly, if a subject serially contests decisions, the deployer might get overloaded with verification tasks, while if a subject serially accepts decisions without contestation, it might lead to an increasing prevalence of inappropriate decisions and eventually shape long-term consequences [165].

Researchers have already started investigating the relation between the decision-subject and the AI system or the combination of the AI system and its user. They have explored the factors that shape such a relation, whether the system is trusted, whether it is deemed fair, and under which conditions [166–168]. Others have explored how to facilitate access to relevant information about the AI system and its decisions for the decision-subject, for them to better determine when to accept a decision [169, 170]. This might go through appropriate information disclosures by the AI deployers or users (e.g., as suggested by certain AI regulations and policies) [171–173], by the development of explanations appropriate for the comprehension of the decision-subjects [174], and by the definition of AI literacy and its development among all kinds of decision-subjects [67, 68]. Finally, prior work has also investigated when and how the subjects might want to contest a decision and what might be technical and procedural structures to facilitate their efforts [175–179].

These insights are particularly relevant to AI deployers who should not only focus on serving a high-performance AI system but also on designing interactions with users and decision-subjects that enable scrutiny, the establishment of appropriate trust, and contestation when relevant. Note that prior work has shown that

without such careful designs, the decision-subjects (or even the AI users) might not consider it fair and might want to contest an AI system and/or the resulting decision [168].

7.4.2 *Upstream AI Users: Broadening the Lens of Human-AI Collaboration to Realities of the AI Supply Chain*

Recent scholarship proposes to expand the scope through which AI is scrutinized. Instead of focusing on the algorithm underlying the AI system, the outputs it produces, and how intended users interact with those outputs, HCI, STS, and legal scholars increasingly draw attention to the *AI supply chain* [180–183] (also referred to as *AI value chain* [184]). The AI supply chain should also be considered as a consequential site of human-AI collaboration, which impacts what the AI system can do and the potential harm it might cause.

7.4.2.1 What Does the AI Supply Chain Constitute?

With the AI supply chain, scholars refer to the various stakeholders who are involved in building, deploying, governing, and maintaining an AI system as well as its various sub-components. For example, consider a large language model that is often used in combination with one or multiple output filters and input prompt restrictions and that requires datasets and algorithms to function. See Fig. 7.2 for a simple illustration of the AI supply chain. Regulatory documents like the AI Act [172] have also started to adopt this supply chain framing. They particularly emphasize the distinction between various organizations: the *AI provider* is the organization that develops an AI system or possesses an AI system and places it on the market, and the *AI deployer* instead is the organization that adopts an AI system (organizations can hold both roles simultaneously). Note that the users traditionally studied in the context of human-AI collaborations might be employed by the AI deployer or make use of the deployer's system via another contractual relation. For instance, a company might serve an AI system for its *employees* to draft emails faster or might let its *clients* contact an AI-based chatbot for customer-service questions. Some organizations might be both providers and deployers (e.g., OpenAI) while providing access to the *general public* to their AI systems.

Researchers have identified the individual stakeholders involved with these provider and deployer organizations. Naturally, various engineers, data scientists, and AI researchers contribute to building components of the AI system [185, 186]. UX researchers and practitioners investigate which capabilities to enable with such a system and how to design it in order to enable these capabilities [187–189]. Additionally, product managers, responsible AI-specific roles, internal lawyers, and policymakers, all bear various roles within these organizations, be it in terms

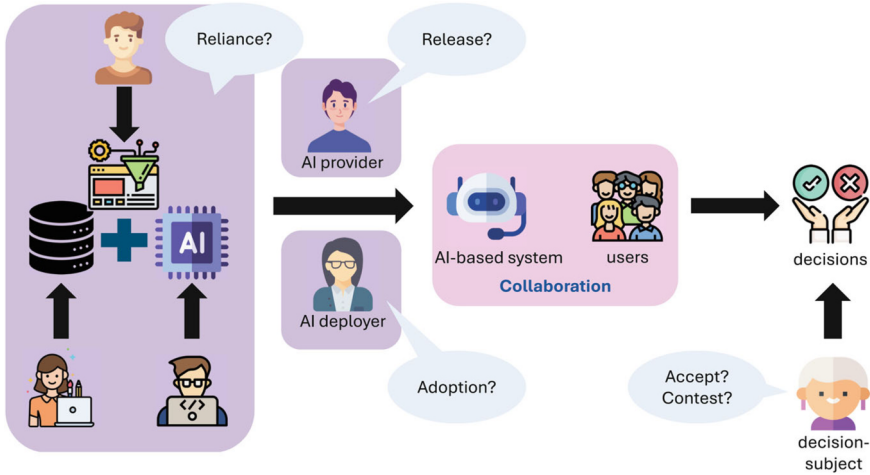


Fig. 7.2 A simplified representation of the sites of human-AI collaboration throughout the lifecycle of an AI-based system. Left of the AI-based system (purple), we show the AI supply chain. Right of the AI-based system (white), we show the decision-subjects. In the speech bubbles (blue), we show examples of questions relevant to the collaboration between each stakeholder and the system or one of its components

of managing teams that produce and release the components or ensuring the compliance with existing regulations (e.g., by identifying potential issues with the systems and setting governance workflows to mitigate these) [183, 190, 191].

Going beyond these central organizations, it is worth noting the materiality of AI supply chains and their resource hunger, particularly in terms of data, labels associated with the data, data storage, computational power, and the required hardware to enable such power [180, 184]. Other stakeholders are typically involved with gathering and developing these resources, such as the crowd workers who annotate datasets or the workers in mines who collect the minerals used to fabricate AI hardware [192–196].

7.4.2.2 Key Notions in the AI Supply Chain

Prior research has addressed two primary knowledge gaps in relation to the AI supply chain—factors that shape the AI supply chain and harms that stem from it. We discuss this next.

Factors That Shape the AI Supply Chain Many studies have surfaced the practices of individual stakeholders in the supply chain and identified the various factors that affect these practices. They have pointed out the individual factors that affect the perceptions and efforts of these stakeholders, such as their knowledge of AI techniques and of AI harms, their reflexivity over their own thoughts and

practices, their socio-demographic and educational background, and their prior experiences with AI and potential harms [185, 186, 197].

Researchers have also identified many organizational challenges that stakeholders face in the supply chain and that typically prevent them from ensuring the development of responsible AI systems, largely due to misaligned interests. They have pointed out the lack of incentives, time, and budget provided to AI developers to test the systems for potential harms, the lack of prioritization of responsible AI efforts in comparison to AI development and release efforts, and the lack of support for educating AI practitioners about the harms of AI [186, 198, 199]. As a result, they have started to propose new governance processes at the organizational level in order to ensure appropriate support for AI practitioners and foster effective collaborations within AI providers or in between AI providers and deployers, to ultimately make sure responsible systems are deployed [200].

Although still rare, some scholarship has also started to focus specifically on the interactions between the stakeholders of the supply chain and how they affect the resulting AI systems. Researchers have also studied the diversity of trust relations that make up the supply chain. In that regard, they have highlighted that many stakeholders form a relation of trust with the AI system or one of its components but that the factors affecting such trust relations might differ from the factors affecting a user's trust in the AI system [183]. They have also shown how trust relations in between organizations or within the stakeholders of a same organization might further affect the development of the final AI system. Others have investigated how responsibility is perceived among these stakeholders and how AI practitioners often perceive it as a concern and obligation for the stakeholders upstream or downstream the supply chain but not for themselves [181, 201]. Note that neither the notion of the AI supply chain nor the association of these various factors to the supply chain is new. Earlier literature in organizational management and other disciplines have conceptualized the notion of supply chains and demonstrated the formative power of these factors [184]. Nevertheless, recent adaptations serve to transfer these notions to the AI context and surface its nuances. For instance, the kind of information that an AI developer requires to develop trust in an AI component such as a dataset is arguably different from the information that a car manufacturer would rely on to establish trust in the car's anti-lock braking system (ABS).

Thus, following these potential collaboration-related challenges that make up AI supply chains, researchers have started to explore the AI-centered information and knowledge that should be shared and communicated among the stakeholders of the supply chain. Transparency and literacy could facilitate the appropriate development and governance of the system [190, 202] if they are enacted appropriately. They might particularly require adaptations depending on the stakeholders involved in the information exchanges, their diverse backgrounds, and their economic relationships (e.g., whether they belong to the same organization, in which case confidentiality and sensitivity of the information might be less of a hindrance for the exchange) [190, 203]. Such communications are also supported by the AI Act [172] that enforces various types of disclosures in between different stakeholders.

Harms That Stem from the AI Supply Chain The second angle that researchers have adopted to investigate AI supply chains is their relation to the notion of AI harms. On the one hand, they have shown that the supply chain might be a source of harm. All the actions taken (or not) by the stakeholders of the supply chain, and all the factors and challenges above, will ultimately affect how biased or fair an AI system is and how transparent it is. They have highlighted that because of supply chain complexities, it is challenging to ensure fairness ultimately [181, 182, 184, 204]. On the other hand, they have discussed how the supply chain itself might be a harmful space. Certain scholars have discussed the poor labor conditions in which certain practitioners work within the supply chain [192]. Other scholars have revealed the power relations among the organizations of the supply chain, the dependencies that are created among these organizations, and the economic and political risk for the organizations with less leverage [182, 204–206].

7.4.2.3 Opportunities to Foster Responsible AI Supply Chains

Despite the little attention that AI supply chains have received, early insights about the stakeholders and dynamics that make up AI supply chains (described above) motivate the importance of their study. These insights particularly demonstrate many challenges in ensuring the supply chain's compliance with regulations, decreasing its potential for harm, and mitigating the issues with the final AI systems it produces. In this context, AI supply chains are very relevant to this chapter. They constitute a new space of human-AI collaborations upstream the usage of the system's outputs by its intended individual users. Indeed, all the stakeholders we mentioned above have to make decisions related to the AI system or its sub-components. A product manager in a deployer organization might, for instance, have to decide whether to adopt the AI system of an AI provider, accounting for various factors such as the reliability of the provider, the trustworthiness of its AI system, and the utility of employing such a system in relation to its cost. Instead, a product manager at a provider organization might have to decide whether to release an AI system and through which modalities (e.g., open-source, via an API access), considering the performance of the system as well as how risky it might be (e.g., what its propensity for outputting offensive text might be and how this might affect the reputation of the provider). Furthermore, all these junctions where a decision is made about the AI system or one of its components end up affecting the final AI system. Such resulting AI systems have been the object of much of human-AI collaboration research. We can appreciate that the properties and attributes of AI systems are shaped all along the supply chain. Hence, better understanding the supply chain also enables better understanding of the collaboration between the AI system and its end-users.

This consideration opens up many research opportunities. It brings to question whether existing approaches developed to support end-user collaborations with AI systems and to study the factors that shape outcomes in such contexts can be transferred to other stakeholders in the supply chain. For instance, it would

be interesting to ask whether the information that users need to learn about to appropriately rely on an AI system are the same as those that AI practitioners (e.g., AI developers, product managers) in the chain would need to make appropriate decisions. It is easy to envision that some of the work will require strong adaptations. In that case, it will be especially important for enterprises to actively consider their role in the supply chain and how to leverage existing research insights to facilitate the appropriate decision-making of each stakeholder vis-a-vis the AI system and its components.

7.4.3 The Generative AI Opportunity Cost for Enterprises Under the Lens of Trust and Reliance

The opportunity costs of investing in generative AI technologies represents a critical consideration for enterprises, with important second-order effects at stake (e.g., on energy consumption and sustainability) [207, 208]. While grappling with the potential and the hype surrounding generative AI, carefully evaluating the opportunity costs and considering the technological trade-offs is vital. We identify some key challenges and trade-offs in this vein, particularly in view of trust and reliance on generative AI in the enterprise context.

- Trust calibration can be a moving target with generative AI. Enterprises should seek to develop meaningful governance structures that cater for the measurement, fostering, and monitoring of appropriate trust along with feedback mechanisms to help calibrate trust [48].
- It is important to create adequate guardrails to help navigate the risks of over-reliance on generative AI. Failing to implement the necessary guardrails can lead to unforeseen costs and harms. For instance, LLMs are prone to hallucinations and inaccuracies that can lead to data pollution when incorporated into enterprise knowledge bases without appropriate verification or validation [209].
- Over-reliance on generative AI can also lead to the dilution of knowledge within enterprises that is typically present in the unstructured mental models of actors and stakeholders in the enterprise, in their routines, and in particular the tacit knowledge of experts [210].
- Over-reliance on generative AI can lead to a decrease in critical thinking [211, 212]. The enterprise opportunity cost manifests here in the form of diminished problem-solving capacity, especially for novel challenges that require creative thinking that is still beyond current AI capabilities.

7.5 Conclusions

There is a growing wealth of knowledge around how we can design and support effective human-AI collaboration. In this chapter, we unpacked the fundamental factors that shape experiential and performance-related outcomes in human-AI collaboration. We discussed the challenges pertaining to fostering appropriate reliance on AI assistance and promising directions stemming from recent literature in this field. We argued that the AI supply chain can be both a valuable lens and a differentiator in informing the effective design and deployment of enterprise AI. An important aspect to consider in the drive toward adoption of generative AI technologies into enterprise practices is that competitive advantages and operational efficiency gains can only emerge from their meaningful and strategic application rather than mere adoption alone. Aligning AI capabilities with the core business objectives; simultaneously considering different stakeholders and customer needs and their values; and explicitly considering second-order effects and societal impact can be a rewarding approach.

Acknowledgments This work was partially supported by the TU Delft AI Initiative, the ProtectMe Convergence Flagship project, and the Model-driven Decisions Lab (MoDDL).

References

1. Licklider, J.C.R.: Man-computer symbiosis. *IRE Trans. Hum. Factors Electron.* **HFE-1**(1), 4–11 (1960). <https://doi.org/10.1109/THFE2.1960.4503259>
2. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al.: Highly accurate protein structure prediction with alphafold. *Nature* **596**(7873), 583–589 (2021)
3. Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A.J., Bambrick, J., et al.: Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**(8016), 493–500 (2024)
4. Shen, J., Zhang, C.J., Jiang, B., Chen, J., Song, J., Liu, Z., He, Z., Wong, S.Y., Fang, P.H. Ming, W.K., et al.: Artificial intelligence versus clinicians in disease diagnosis: systematic review. *JMIR Med. Inf.* **7**(3), e10010 (2019)
5. Saab, K., Tu, T., Weng, W.H., Tanno, R., Stutz, D., Wulczyn, E., Zhang, F., Strother, T., Park, C., Vedadi, E., et al.: Capabilities of gemini models in medicine. Preprint. arXiv:2404.18416 (2024)
6. Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., et al.: Towards expert-level medical question answering with large language models. Preprint. arXiv:2305.09617 (2023)
7. Uddin, M., Wang, Y., Woodbury-Smith, M.: Artificial intelligence for precision medicine in neurodevelopmental disorders. *npj Digital Med.* **2**(1), 112 (2019)
8. Eyring, V., Gentine, P., Camps-Valls, G., Lawrence, D.M., Reichstein, M.: AI-empowered next-generation multiscale climate modelling for mitigation and adaptation. *Nat. Geosci.* **17**(10), 963–971 (2024)
9. Silvestro, D., Gorla, S., Sterner, T., Antonelli, A.: Improving biodiversity protection through artificial intelligence. *Nat. Sustainability* **5**(5), 415–424 (2022)

10. Følstad, A., Skjuve, M.: Chatbots for customer service: user experience and motivation. In: Proceedings of the 1st International Conference on Conversational User Interfaces, pp. 1–9 (2019)
11. Moilanen, J., van Berkel, N., Visuri, A., Gadiraju, U., van der Maden, W., Hosio, S.: Supporting mental health self-care discovery through a chatbot. *Front. Digital Health* **5**, 1034724 (2023)
12. Abbas, T., Khan, V.J., Gadiraju, U., Markopoulos, P.: Trainbot: a conversational interface to train crowd workers for delivering on-demand therapy. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, vol. 8, pp. 3–12 (2020)
13. Li, B., Gross, O., Crampton, N., Kapoor, M., Tauseef, S., Jain, M., Truong, K.N., Mariakakis, A.: Beyond the waiting room: patient’s perspectives on the conversational nuances of pre-consultation chatbots. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 1–24 (2024)
14. Kuhail, M.A., Alturki, N., Alramlawi, S., Alhejori, K.: Interacting with educational chatbots: A systematic review. *Educ. Inf. Technol.* **28**(1), 973–1018 (2023)
15. McGregor, S.: Preventing repeated real world AI failures by cataloging incidents: the AI incident database. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 15458–15463 (2021)
16. Constantinides, M., Bogucka, E., Quercia, D., Kallio, S., Tahaei, M.: RAI guidelines: method for generating responsible AI guidelines grounded in regulations and usable by (non-) technical roles. *Proc. ACM Hum.-Comput. Interact.* **8**(CSCW2), 1–28 (2024)
17. Shneiderman, B.: Human-centered artificial intelligence: Reliable, safe & trustworthy. *Int. J. Hum.-Comput. Interact.* **36**(6), 495–504 (2020)
18. Cavalcante Siebert, L., Lupetti, M.L., Aizenberg, E., Beckers, N., Zgonnikov, A., Veluwenkamp, H., Abbink, D., Giaccardi, E., Houben, G.J., Jonker, C.M., et al.: Meaningful human control: actionable properties for ai system development. *AI Ethics* **3**(1), 241–255 (2023)
19. Amershi, S., Weld, D., Vorvoreanu, M., Fournay, A., Nushi, B., Collis, P., Suh, J., Iqbal, S., Bennett, P.N., Inkpen, K., et al.: Guidelines for human-AI interaction. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–13 (2019)
20. Gadiraju, U., Yang, J.: What can crowd computing do for the next generation of AI systems? In: *CSW@ NeurIPS*, pp. 7–13 (2020)
21. De Cremer, D., Kasparov, G.: AI should augment human intelligence, not replace it. *Harv. Bus. Rev.* **18**(1), 1–8 (2021)
22. Wang, D., Churchill, E., Maes, P., Fan, X., Shneiderman, B., Shi, Y., Wang, Q.: From human-human collaboration to Human-AI collaboration: designing AI systems that can work together with people. In: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–6 (2020)
23. Fan, M., Yang, X., Yu, T., Liao, Q.V., Zhao, J.: Human-AI collaboration for UX evaluation: effects of explanation and synchronization. *Proc. ACM Hum.-Comput. Interact.* **6**(CSCW1), 1–32 (2022)
24. Vodrahalli, K., Gerstenberg, T., Zou, J.Y., Uncalibrated models can improve human-AI collaboration. *Adv. Neural Inf. Proces. Syst.* **35**, 4004–4016 (2022)
25. Reverberi, C., Rigon, T., Solari, A., Hassan, C., Cherubini, P., Cherubini, A.: Experimental evidence of effective human–AI collaboration in medical decision-making. *Sci. Rep.* **12**(1), 14952 (2022)
26. Hemmer, P., Westphal, M., Schemmer, M., Vetter, S., Vössing, M., Satzger, G.: Human-AI collaboration: the effect of AI delegation on human task performance and task satisfaction. In: Proceedings of the 28th International Conference on Intelligent User Interfaces, pp. 453–463 (2023)
27. Ashktorab, Z., Liao, Q.V., Dugan, C., Johnson, J., Pan, Q., Zhang, W., Kumaravel, S., Campbell, M.: Human-AI collaboration in a cooperative game setting: measuring social perception and outcomes. *Proc. ACM Hum.-Comput. Interact.* **4**(CSCW2), 1–20 (2020)

28. Sowa, K., Przegalinska, A., Ciechanowski, L.: Cobots in knowledge work: Human–AI collaboration in managerial professions. *J. Bus. Res.* **125**, 135–142 (2021)
29. Wang, D., Weisz, J.D., Muller, M., Ram, P., Geyer, W., Dugan, C., Tausczik, Y., Samulowitz, H., Gray, A.: Human-AI collaboration in data science: exploring data scientists’ perceptions of automated AI. *Proc. ACM Hum.-Comput. Interact.* **3**(CSCW), 1–24 (2019)
30. Jiang, J.A., Wade, K., Fiesler, C., Brubaker, J.R.: Supporting serendipity: opportunities and challenges for human-ai collaboration in qualitative analysis. *Proc. ACM Hum.-Comput. Interact.* **5**(CSCW1), 1–23 (2021)
31. Rastogi, C., Tulio Ribeiro, M., King, N., Nori, H., Amershi, S.: Supporting human-AI collaboration in auditing LLMs with LLMs. In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 913–926 (2023)
32. McNeese, N.J., Demir, M., Cooke, N.J., Myers, C.: Teaming with a synthetic teammate: insights into human-autonomy teaming. *Hum. Fact.* **60**(2), 262–273 (2018)
33. O’Neill, T., McNeese, N., Barron, A., Schelble, B.: Human–autonomy teaming: a review and analysis of the empirical literature. *Hum. Fact.* **64**(5), 904–938 (2022)
34. Seeber, I., Bittner, E., Briggs, R.O., De Vreede, T., De Vreede, G.J., Elkins, A., Maier, R., Merz, A.B., Oeste-Reiß, S., Randrup, N., et al.: Machines as teammates: a research agenda on AI in team collaboration. *Inf. Manag.* **57**(2), 103174 (2020)
35. Hauptman, A.I., Schelble, B.G., McNeese, N.J., Madathil, K.C.: Adapt and overcome: perceptions of adaptive autonomous agents for human-AI teaming. *Comput. Hum. Behav.* **138**, 107451 (2023)
36. Demartini, G., Difallah, D.E., Gadiraju, U., Catasta, M., et al.: An introduction to hybrid human-machine information systems. *Found. Trends® Web Sci.* **7**(1), 1–87 (2017)
37. Dell’Anna, D., Murukannaiah, P.K., Dudzik, B., Grossi, D., Jonker, C.M., Oertel, C., Yolum, P.: Toward a quality model for hybrid intelligence teams. In: *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pp. 434–443 (2024)
38. Zhang, R., McNeese, N.J., Freeman, G., Musick, G.: “an ideal human” expectations of AI teammates in human-AI teaming. *Proc. ACM Hum.-Comput. Interact.* **4**(CSCW3), 1–25 (2021)
39. Zhang, G., Chong, L., Kotovsky, K., Cagan, J.: Trust in an AI versus a human teammate: the effects of teammate identity and performance on human-AI cooperation. *Comput. Hum. Behav.* **139**, 107536 (2023)
40. He, G., Demartini, G., Gadiraju, U.: Plan-then-execute: an empirical study of user trust and team performance when using LLM agents as a daily assistant. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2025)
41. He, G., Aishwarya, N., Gadiraju, U.: Is conversational XAI all you need? Human-AI decision making with a conversational XAI assistant. In: *Proceedings of the 30th International Conference on Intelligent User Interfaces* (2025)
42. Kim, S.S., Liao, Q.V., Vorvoreanu, M., Ballard, S., Vaughan, J.W.: “I’m Not Sure, But . . .”: examining the impact of large language models’ uncertainty expression on user reliance and trust. In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 822–835 (2024)
43. Vasconcelos, H., Bansal, G., Fourney, A., Liao, Q.V., Vaughan, J.W.: Generation probabilities are not enough: exploring the effectiveness of uncertainty highlighting in ai-powered code completions. Preprint. arXiv:2302.07248 (2023)
44. Hechler, E., Oberhofer, M., Schaeck, T.: Deploying ai in the enterprise. In: *IT Approaches for Design, DevOps, Governance, Change Management, Blockchain, and Quantum Computing*, Apress, Berkeley, CA (2020)
45. Lockey, S., Gillespie, N., Morrill, J., Pool, J.: Trust in artificial intelligence: a critical systematic review. In: *Academy of Management Proceedings*, vol. 2024, p. 14990. Academy of Management, Valhalla (2024)
46. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* **267**, 1–38 (2019)

47. Lee, J.D., See, K.A.: Trust in automation: designing for appropriate reliance. *Hum. Fact.* **46**(1), 50–80 (2004)
48. Hoff, K.A., Bashir, M.: Trust in automation: Integrating empirical evidence on factors that influence trust. *Hum. Fact.* **57**(3), 407–434 (2015)
49. Mehrotra, S., Degachi, C., Vereschak, O., Jonker, C.M., Tielman, M.L.: A systematic review on fostering appropriate trust in human-AI interaction: trends, opportunities and challenges. *ACM J. Responsible Comput.* **1**(4), 1–45 (2024)
50. Schlicker, N., Langer, M.: Towards warranted trust: a model on the relation between actual and perceived system trustworthiness. In: *Proceedings of Mensch und Computer 2021*, pp. 325–329 (2021)
51. Akula, A., Wang, S., Zhu, S.C.: Cocox: generating conceptual and counterfactual explanations via fault-lines. In: *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, pp. 2594–2601 (2020)
52. Jacovi, A., Marasović, A., Miller, T., Goldberg, Y.: Formalizing trust in artificial intelligence: prerequisites, causes and goals of human trust in AI. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 624–635 (2021)
53. Ferrario, A., Loi, M.: How explainability contributes to trust in AI. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1457–1466 (2022)
54. Lebiere, C., Blaha, L.M., Fallon, C.K., Jefferson, B.: Adaptive cognitive mechanisms to maintain calibrated trust and reliance in automation. *Front. Rob. AI* **8**, 652776 (2021)
55. Ghai, B., Liao, Q.V., Zhang, Y., Bellamy, R., Mueller, K.: Explainable active learning (XAL) toward AI explanations as interfaces for machine teachers. *Proc. ACM Hum.-Comput. Interact.* **4**(CSCW3), 1–28 (2021)
56. Schlicker, N., Baum, K., Uhde, A., Sterz, S., Hirsch, M.C., Langer, M.: How do we assess the trustworthiness of AI? Introducing the trustworthiness assessment model (TrAM). *Comput. Hum. Behav.* **170**, 108671 (2025)
57. Miller, T.: Are we measuring trust correctly in explainability, interpretability, and transparency research? Preprint. arXiv:2209.00651 (2022)
58. Okamura, K., Yamada, S.: Adaptive trust calibration for human-ai collaboration. *Plos One* **15**(2), e0229132 (2020)
59. Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M.T., Weld, D.: Does the whole exceed its parts? The effect of ai explanations on complementary team performance. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–16 (2021)
60. Inkpen, K., Chappidi, S., Mallari, K., Nushi, B., Ramesh, D., Michelucci, P., Mandava, V., Vepřek, L.H., Quinn, G.: Advancing human-ai complementarity: the impact of user expertise and algorithmic tuning on joint decision making. *ACM Trans. Comput.-Hum. Interact.* **30**(5), 1–29 (2023)
61. Lai, V., Chen, C., Liao, Q.V., Smith-Renner, A., Tan, C.: Towards a science of human-AI decision making: a survey of empirical studies. Preprint. arXiv:2112.11471 (2021)
62. Schemmer, M., Hemmer, P., Kühl, N., Benz, C., Satzger, G.: Should I follow ai-based advice? Measuring appropriate reliance in human-ai decision-making. Preprint. arXiv:2204.06916 (2022)
63. Salimzadeh, S., He, G., Gadiraju, U.: Dealing with uncertainty: understanding the impact of prognostic versus diagnostic tasks on trust and reliance in human-AI decision making. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–17 (2024)
64. Erlei, A., Das, R., Meub, L., Anand, A., Gadiraju, U.: For what it's worth: humans overwrite their economic self-interest to avoid bargaining with AI systems. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–18 (2022)
65. Hemmer, P., Schemmer, M., Vössing, M., Kühl, N.: Human-AI complementarity in hybrid intelligence systems: a structured literature review. *PACIS* **78**, 118 (2021)

66. Hartikainen, M., Spurava, G., Väänänen, K.: Human-AI collaboration in smart manufacturing: key concepts and framework for design. *HHAI 2024: Hybrid Hum. AI Syst. Soc. Good* 162–172 (2024)
67. Long, D., Magerko, B.: What is AI literacy? Competencies and design considerations. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–16 (2020)
68. NDg, D.T.K., LJeung, J.K.L., CShu, S.K.W., Qiao, M.S.: Conceptualizing ai literacy: an exploratory review. *Comput. Educ.: Artif. Intell.* **2**, 100041 (2021)
69. Pinski, M., Benlian, A.: AI literacy for users-a comprehensive review and future research directions of learning methods, components, and effects. *Comput. Hum. Behav.: Artif. Hum.* **2**(1), 100062 (2024)
70. Salimzadeh, S., He, G., Gadiraju, U.: A missing piece in the puzzle: considering the role of task complexity in human-ai decision making. In: *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, pp. 215–227 (2023)
71. Nourani, M., King, J., Ragan, E.: The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 8, pp. 112–121 (2020)
72. Senoner, J., Schallmoser, S., Kratzwald, B., Feuerriegel, S., Netland, T.: Explainable AI improves task performance in human–AI collaboration. *Sci. Rep.* **14**(1), 31150 (2024)
73. Zhang, Q., Lee, M.L., Carter, S.: You complete me: human-AI teams and complementary expertise. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–28 (2022)
74. Baron, S., Latham, A.J., Varga, S.: Explainable ai and stakes in medicine: a user study. *Artif. Intell.* **340**, 104282 (2025)
75. Yin, M., Wortman Vaughan, J., Wallach, H.: Understanding the effect of accuracy on trust in machine learning models. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–12 (2019)
76. He, G., Buijsman, S., Gadiraju, U.: How stated accuracy of an ai system and analogies to explain accuracy affect human reliance on the system. *Proc. ACM Hum.-Comput. Interact.* **7**(CSCW2), 1–29 (2023)
77. Bansal, G., Nushi, B., Kamar, E., Lasecki, W.S., Weld, D.S., Horvitz, E.: Beyond accuracy: the role of mental models in human-AI team performance. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 7, pp. 2–11 (2019)
78. Fan, X., Oh, S., McNeese, M., Yen, J., Cuevas, H., Strater, L., Endsley, M.R.: The influence of agent reliability on trust in human-agent collaboration. In: *Proceedings of the 15th European Conference on Cognitive Ergonomics: The Ergonomics of Cool Interaction*, pp. 1–8 (2008)
79. Appalganc, K., Rieger, T., Roesler, E., Manzey, D.: How much reliability is enough? A context-specific view on human interaction with (artificial) agents from different perspectives. *J. Cogn. Eng. Decis. Mak.* **16**(4), 207–221 (2022). <https://doi.org/10.1177/15553434221104615>
80. Nourani, M., Roy, C., Block, J.E., Honeycutt, D.R., Rahman, T., Ragan, E., Gogate, V.: Anchoring bias affects mental model formation and user reliance in explainable AI systems. In: *Proceedings of the 26th International Conference on Intelligent User Interfaces*, pp. 340–350 (2021)
81. Tolmeijer, S., Gadiraju, U., Ghantasala, R., Gupta, A., Bernstein, A.: Second chance for a first impression? Trust development in intelligent system interaction. In: *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, pp. 77–87 (2021)
82. Swaroop, S., Bućinca, Z., Gajos, K.Z., Doshi-Velez, F.: Accuracy-time tradeoffs in AI-assisted decision making under time pressure. In: *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pp. 138–154 (2024)
83. Cao, S., Gomez, C., Huang, C.M.: How time pressure in different phases of decision-making influences human-AI collaboration. *Proc. ACM Hum.-Comput. Interact.* **7**(CSCW2), 1–26 (2023)

84. He, G., Kuiper, L., Gadiraju, U.: Knowing about knowing: an illusion of human competence can hinder appropriate reliance on AI systems. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–18 (2023)
85. Khadpe, P., Krishna, R., Fei-Fei, L., Hancock, J.T., Bernstein, M.S.: Conceptual metaphors impact perceptions of human-AI collaboration. *Proc. ACM Hum.-Comput. Interact.* **4**(CSCW2), 1–26 (2020)
86. Jung, J.Y., Qiu, S., Bozzon, A., Gadiraju, U.: Great chain of agents: the role of metaphorical representation of agents in conversational crowdsourcing. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–22 (2022)
87. Szymanski, M., Millicamp, M., Verbert, K.: Visual, textual or hybrid: the effect of user expertise on different explanations. In: *Proceedings of the 26th International Conference on Intelligent User Interfaces*, pp. 109–119 (2021)
88. Robbemon, V., Inel, O., Gadiraju, U.: Understanding the role of explanation modality in AI-assisted decision-making. In: *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, pp. 223–233 (2022)
89. Van de Poel, I.: Embedding values in artificial intelligence (AI) systems. *Minds Mach.* **30**(3), 385–409 (2020)
90. Arora, P.: *From Pessimism to Promise: Lessons from the Global South on Designing Inclusive Tech*. MIT Press, Cambridge (2024)
91. Cetindamar, D., Kitto, K., Wu, M., Zhang, Y., Abedin, B., Knight, S.: Explicating AI literacy of employees at digital workplaces. *IEEE Trans. Eng. Manage.* **71**, 810–823 (2022)
92. Carolus, A., Koch, M.J., Straka, S., Latoschik, M.E., Wienrich, C.: Mails-meta AI literacy scale: development and testing of an ai literacy questionnaire based on well-founded competency models and psychological change-and meta-competencies. *Comput. Hum. Behav. Artif. Hum.* **1**(2), 100014 (2023)
93. Lintner, T.: A systematic review of ai literacy scales. *NPJ Sci. Learn.* **9**(1), 50 (2024)
94. Kasinidou, M.: Promoting AI literacy for the public. In: *Proceedings of the 54th ACM Technical Symposium on Computer Science Education*, vol. 2, pp. 1237–1237 (2023)
95. Su, J., NDg, D.T.K., CShu, S.K.W.: Artificial intelligence (AI) literacy in early childhood education: the challenges and opportunities. *Comput. Educ. Artif. Intell.* **4**, 100124 (2023)
96. NDg, D.T.K., Xinyu, C., LJeung, J.K.L., CShu, S.K.W.: Fostering students' AI literacy development through educational games: AI knowledge, affective and cognitive engagement. *J. Comput. Assist. Learn.* **40**(5), 2049–2064 (2024)
97. Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J.: Measures for explainable AI: explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Front. Comput. Sci.* **5**, 1096257 (2023)
98. Andrews, R.W., Lilly, J.M., Srivastava, D., Feigh, K.M.: The role of shared mental models in human-AI teams: a theoretical review. *Theor. Issues Ergonom. Sci.* **24**(2), 129–175 (2023)
99. Gero, K.I., Ashktorab, Z., Dugan, C., Pan, Q., Johnson, J., Geyer, W., Ruiz, M., Miller, S., Millen, D.R., Campbell, M., et al.: Mental models of AI agents in a cooperative game setting. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–12 (2020)
100. Cabrera, Á.A., Perer, A., Hong, J.I.: Improving human-AI collaboration with descriptions of AI behavior. *Proc. ACM Hum.-Comput. Interact.* **7**(CSCW1), 1–21 (2023)
101. Nourani, M., Roy, C., Block, J.E., Honeycutt, D.R., Rahman, T., Ragan, E.D., Gogate, V.: On the importance of user backgrounds and impressions: lessons learned from interactive ai applications. *ACM Trans. Interact. Intell. Syst.* **12**(4), 1–29 (2022)
102. Nourani, M., Honeycutt, D.R., Block, J.E., Roy, C., Rahman, T., Ragan, E.D., Gogate, V.: Investigating the importance of first impressions and explainable AI with interactive video analysis. In: *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–8 (2020)
103. Wilder, B., Horvitz, E., Kamar, E.: Learning to complement humans. In: *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pp. 1526–1533 (2021)

104. Kahneman, D.: *Thinking, Fast and Slow* (Macmillan, New York, 2011)
105. Bertrand, A., Belloum, R., Eagan, J.R., Maxwell, W.: How cognitive biases affect XAI-assisted decision-making: a systematic review. In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 78–91 (2022)
106. Echterhoff, J.M., Yarmand, M., McAuley, J.: AI-moderated decision-making: capturing and balancing anchoring bias in sequential decision tasks. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–9 (2022)
107. Rastogi, C., Zhang, Y., Wei, D., Varshney, K.R., Dhurandhar, A., Tomsett, R.: Deciding fast and slow: the role of cognitive biases in AI-assisted decision-making. *Proc. ACM Hum.-Comput. Interact.* **6**(CSCW1), 1–22 (2022)
108. Jacobsen, R.M., BLysted, L.B.L., Johansen, P.S., Papachristos, E., Skov, M.B.: Perceived and measured task effectiveness in human-AI collaboration. In: *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–9 (2020)
109. Biswas, S., Erlei, A., Gadiraju, U.: Mind the gap! choice independence in using multilingual LLMs for persuasive co-writing tasks in different languages. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2025)
110. Sharifi Noorian, S., Qiu, S., Sayin, B., Balayn, A., Gadiraju, U., Yang, J., Bozzon, A.: Perspective: leveraging human understanding for identifying and characterizing image atypicality. In: *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pp. 650–663 (2023)
111. Zhang, Z., Singh, J., Gadiraju, U., Anand, A.: Dissonance between human and machine understanding. *Proc. ACM Hum.-Comput. Interact.* **3**(CSCW), 1–23 (2019)
112. Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., Steinhardt, J.: Aligning AI with shared human values. In: *International Conference on Learning Representations*
113. Seth, A., Ahuja, S., Bali, K., Sitaram, S.: DOSA: a dataset of social artifacts from different indian geographical subcultures. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 5323–5337 (2024)
114. Christian, B.: *The Alignment Problem: How Can Machines Learn Human Values?* Atlantic Books, London (2021)
115. Lai, V., Chen, C., Smith-Renner, A., Liao, Q.V., Tan, C.: Towards a science of human-AI decision making: an overview of design space in empirical human-subject studies. In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1369–1385 (2023)
116. Salimzadeh, S., Gadiraju, U.: When in doubt! understanding the role of task characteristics on peer decision-making with AI assistance. In: *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, pp. 89–101 (2024)
117. Tomsett, R., Preece, A., Braines, D., Cerutti, F., Chakraborty, S., Srivastava, M., Pearson, G., Kaplan, L.: Rapid trust calibration through interpretable and uncertainty-aware AI. *Patterns* **1**(4), 100049 (2020). ISSN: 2666–3899. <https://doi.org/10.1016/j.patter.2020.100049>
118. Prabhudesai, S., Yang, L., Asthana, S., Huan, X., Liao, Q.V., Banovic, N.: Understanding uncertainty: how lay decision-makers perceive and interpret uncertainty in human-AI decision making. In: *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pp. 379–396 (2023)
119. Melis, D.A., Kaur, H., Daumé III, H., Wallach, H., Vaughan, J.W.: From human explanation to model interpretability: a framework based on weight of evidence. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 9, pp. 35–47 (2021)
120. Rechkemmer, A., Yin, M.: When confidence meets accuracy: exploring the effects of multiple performance indicators on trust in machine learning models. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–14 (2022)
121. Ahn, D., Almaatouq, A., Gulabani, M., Hosanagar, K.: Impact of model interpretability and outcome feedback on trust in AI. In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–25 (2024)

122. Kaur, H., Conrad, M.R., Rule, D., Lampe, C., Gilbert, E.: Interpretability gone bad: the role of bounded rationality in how practitioners understand machine learning. *Proc. ACM Hum.-Comput. Interact.* **8**(CSCW1), 1–34 (2024)
123. Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., Wortman Vaughan, J.: Interpreting interpretability: understanding data scientists’ use of interpretability tools for machine learning. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–14 (2020)
124. Bansal, G., Nushi, B., Kamar, E., Weld, D.S., Lasecki, W.S., Horvitz, E.: Updates in human-AI teams: understanding and addressing the performance/compatibility tradeoff. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 2429–2437 (2019)
125. Zhang, R., Duan, W., Flathmann, C., McNeese, N., Knijnenburg, B., Freeman, G.: Verbal vs. visual: how humans perceive and collaborate with AI teammates using different communication modalities in various human-AI team compositions. *Proc. ACM Hum.-Comput. Interact.* **8**(CSCW2), 1–34 (2024)
126. Liu, N., Pu, Q., Shi, Y., Zhang, S., Qiu, L.: Older adults’ interaction with intelligent virtual assistants: the role of information modality and feedback. *Int. J. Hum.-Comput. Interact.* **39**(5), 1162–1183 (2023)
127. Le Bigot, L., Terrier, P., Amiel, V., Poulain, G., Jamet, E., Rouet, J.F.: Effect of modality on collaboration with a dialogue system. *Int. J. Hum.-Comput. Stud.* **65**(12), 983–991 (2007)
128. He, G., Aishwarya, N., Gadiraju, U.: Is conversational XAI all you need? Human-AI decision making with a conversational XAI assistant. In: *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pp. 907–924 (2025)
129. Cai, C.J., Winter, S., Steiner, D., Wilcox, L., Terry, M.: “hello ai”: uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making. *Proc. ACM Hum.-Comput. Interact.* **3**(CSCW), 1–24 (2019)
130. Wang, J., Molina, M.D., Sundar, S.S.: When expert recommendation contradicts peer opinion: relative social influence of valence, group identity and artificial intelligence. *Comput. Hum. Behav.* **107**, 106278 (2020)
131. Grønsund, T., Aanestad, M.: Augmenting the algorithm: emerging human-in-the-loop work configurations. *J. Strateg. Inf. Syst.* **29**(2), 101614 (2020)
132. Wang, B.Y., Boell, S.K., Riemer, K., Peter, S.: Human agency in AI configurations supporting organizational decision-making. In: *ACIS 2023 Proceedings*, vol. 53 (2023). <https://aisel.aisnet.org/acis2023/53>
133. Fogliato, R., Chappidi, S., Lungren, M., Fisher, P., Wilson, D., Fitzke, M., Parkinson, M., Horvitz, E., Inkpen, K., Nushi, B.: Who goes first? influences of human-AI workflow on decision making in clinical imaging. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1362–1374 (2022)
134. Shin, J., Hedderich, M.A., Rey, B.J., Lucero, A., Oulasvirta, A.: Understanding human-AI workflows for generating personas. In: *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, pp. 757–781 (2024)
135. Wu, T., Terry, M., Cai, C.J.: AI chains: transparent and controllable human-AI interaction by chaining large language model prompts. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–22 (2022)
136. Allen, G., He, G., Gadiraju, U.: Power-up! What can generative models do for human computation workflows? Preprint. arXiv:2307.02243 (2023)
137. Wu, T., Zhu, H., Albayrak, M., Axon, A., Bertsch, A., Deng, W., Ding, Z., Guo, B., Gururaja, S., Kuo, T.S., et al.: LLMs as workers in human-computational algorithms? Replicating crowdsourcing pipelines with LLMs. Preprint. arXiv:2307.10168 (2023)
138. Schoeffer, J., De-Arteaga, M., Kuehl, N.: Explanations, fairness, and appropriate reliance in human-AI decision-making. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–18 (2024)
139. He, G., Gadiraju, U.: Walking on eggshells: using analogies to promote appropriate reliance in human-AI decision making. In: *Proceedings of the Workshop on Trust and Reliance on AI-*

- Human Teams at the ACM Conference on Human Factors in Computing Systems (CHI'22) (2022)
140. Küper, A., Krämer, N.: Psychological traits and appropriate reliance: factors shaping trust in AI. *Int. J. Hum.-Comput. Interact.* **41**(7), 4115–4131 (2025)
 141. Guo, Z., Wu, Y., Hartline, J.D., Hullman, J.: A decision theoretic framework for measuring AI reliance. In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 221–236 (2024)
 142. Schlicker, N., Baum, K., Uhde, A., Sterz, S., Hirsch, M.C., Langer, M.: How do we assess the trustworthiness of AI? Introducing the trustworthiness assessment model (TrAM). *Comput. Hum. Behav.* **170**, 108671 (2025)
 143. Schemmer, M., Kuehl, N., Benz, C., Bartos, A., Satzger, G.: Appropriate reliance on AI advice: conceptualization and the effect of explanations. In: *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pp. 410–422 (2023)
 144. Benda, N.C., Novak, L.L., Reale, C., Ancker, J.S.: Trust in AI: why we should be designing for appropriate reliance. *J. Am. Med. Inf. Assoc.* **29**(1), 207–212 (2022)
 145. Kahr, P.K., Rooks, G., Willemsen, M.C., Snijders, C.C.: Understanding trust and reliance development in AI advice: assessing model accuracy, model explanations, and experiences from previous interactions. *ACM Trans. Interact. Intell. Syst.* **14**(4), 1–30 (2024)
 146. Upadhyay, S., Lakkaraju, H., Gajos, K.Z.: Counterfactual explanations may not be the best algorithmic recourse approach. In: *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pp. 446–462 (2025)
 147. Liao, Q.V., Gruen, D., Miller, S.: Questioning the AI: informing design practices for explainable AI user experiences. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–15 (2020)
 148. Liao, Q.V., Varshney, K.R.: Human-centered explainable AI (XAI): from algorithms to user experiences. Preprint. arXiv:2110.10790 (2021)
 149. Liao, Q.V., Pribić, M., Han, J., Miller, S., Sow, D.: Question-driven design process for explainable ai user experiences. Preprint. arXiv:2104.03483 (2021)
 150. Liao, Q.V., Zhang, Y., Luss, R., Doshi-Velez, F., Dhurandhar, A.: Connecting algorithmic research and usage contexts: a perspective of contextualized evaluation for explainable AI. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 10, pp. 147–159 (2022)
 151. Liao, Q.V., Vaughan, J.W.: AI transparency in the age of LLMs: a human-centered research roadmap. Preprint. arXiv:2306.01941, 10 (2023)
 152. Kim, S.S., Vaughan, J.W., Liao, Q.V., Lombrozo, T., Russakovsky, O.: Fostering appropriate reliance on large language models: the role of explanations, sources, and inconsistencies. In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–19 (2025)
 153. Ehsan, U., Liao, Q.V., Passi, S., Riedl, M.O., Daumé III, H.: Seamful xai: operationalizing seamful design in explainable AI. *Proc. ACM Hum.-Comput. Interact.* **8**(CSCW1), 1–29 (2024)
 154. Ehsan, U., Passi, S., Liao, Q.V., Chan, L., Lee, I.H., Muller, M., Riedl, M.O.: The who in XAI: how AI background shapes perceptions of AI explanations. In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–32 (2024)
 155. Schmuckler, M.A.: What is ecological validity? A dimensional analysis. *Infancy* **2**(4), 419–436 (2001)
 156. Kihlstrom, J.F.: Ecological validity and “ecological validity”. *Perspect. Psychol. Sci.* **16**(2), 466–471 (2021)
 157. Fok, R., Weld, D.S.: In search of verifiability: explanations rarely enable complementary performance in ai-advised decision making. *AI Magazine* **45**(3), 317–332 (2024)
 158. Miller, T.: Explainable AI is dead, long live explainable AI! Hypothesis-driven decision support using evaluative AI. In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 333–342 (2023)

159. Buçinca, Z., Malaya, M.B., Gajos, K.Z.: To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proc. ACM Hum.-Comput. Interact.* **5**(CSCW1), 1–21 (2021)
160. He, G., Bharos, A., Gadiraju, U.: To Err Is AI! Debugging as an intervention to facilitate appropriate reliance on AI systems. In: *Proceedings of the 35th ACM Conference on Hypertext and Social Media*, pp. 98–105 (2024)
161. Kahr, P.K., Rooks, G., Willemsen, M.C., Snijders, C.C.: It seems smart, but it acts stupid: development of trust in ai advice in a repeated legal decision-making task. In: *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pp. 528–539 (2023)
162. Kahr, P., Rooks, G., Snijders, C.C., Willemsen, M.C.: Good performance isn't enough to trust AI. Lessons from logistics experts on their long-term collaboration with an AI planning system. In: *ACM CHI 2025 Conference on Human Factors in Computing Systems*. Association for Computing Machinery, Inc, New York (2025)
163. He, G., Hemmer, P., Vössing, M., Schemmer, M., Gadiraju, U.: Fine-grained appropriate reliance: human-AI collaboration with a multi-step transparent decision workflow for complex task decomposition. Preprint. arXiv:2501.10909 (2025)
164. Balayn, A., Pi, Y., Widder, D.G., Alfrink, K., Yurrita, M., Upadhyay, S., Karusala, N., Lyons, H., Turkay, C., Tessono, C., et al.: From stem to stern: contestability along AI value chains. In: *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*, pp. 720–723 (2024)
165. D'Amour, A., Srinivasan, H., Atwood, J., Baljekar, P., Sculley, D., Halpern, Y.: Fairness is not static: deeper understanding of long term fairness via simulation studies. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 525–534 (2020)
166. Yurrita, M., Draws, T., Balayn, A., Murray-Rust, D., Tintarev, N., Bozzon, A.: Disentangling fairness perceptions in algorithmic decision-making: the effects of explanations, human oversight, and contestability. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. Association for Computing Machinery, New York (2023). <https://doi.org/10.1145/3544548.3581161>
167. Yurrita, M., Verma, H., Balayn, A., Gadiraju, U., Pont, S., Bozzon, A.: Towards effective human intervention in algorithmic decision-making: understanding the effect of decision-makers' configuration on decision-subjects' fairness perceptions. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2025)
168. Decker, M.C., Wegner, L., Leicht-Scholten, C.: Procedural fairness in algorithmic decision-making: the role of public engagement. *Ethics Inf. Technol.* **27**(1), 1 (2025)
169. Yurrita, M., Verma, H., Balayn, A., Alfrink, K., Gadiraju, U., Bozzon, A.: Identifying algorithmic decision subjects' needs for meaningful contestability. In: *Proceedings of the CSCW Conference on Computer-Supported Collaborative Work* (2025)
170. Walmsley, J.: Artificial intelligence and the value of transparency. *AI Soc.* **36**(2), 585–595 (2021)
171. Kaminski, M.E., Urban, J.M.: The right to contest AI. *Columbia Law Rev.* **121**(7), 1957–2048 (2021)
172. Madiaga, T.: *Artificial Intelligence Act*. European Parliament: European Parliamentary Research Service, Brussels (2021)
173. Voigt, P., Von dem Bussche, A.: *The EU General Data Protection Regulation (GDPR). A Practical Guide*, 1st edn., Springer International Publishing, Cham (2017)
174. Cheng, H.F., Wang, R., Zhang, Z., O'connell, F., Gray, T., Harper, F.M., Zhu, H.: Explaining decision-making algorithms through UI: strategies to help non-expert stakeholders. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–12 (2019)
175. Alfrink, K., Keller, I., Semperena, M.Y., Bulygin, D., Kortuem, G., Doorn, N.: Envisioning contestability loops: evaluating the agonistic arena as a generative metaphor for public AI. *She Ji: J. Des. Econ. Innovation* **10**(1), 53–93 (2024)

176. Lyons, H., Velloso, E., Miller, T.: Conceptualising contestability: perspectives on contesting algorithmic decisions. *Proc. ACM Hum.-Comput. Interact.* **5**(CSCW1), 1–25 (2021)
177. Alfrink, K., Keller, I., Doorn, N., Kortuem, G.: Contestable camera cars: a speculative design exploration of public AI that is open and responsive to dispute. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–16 (2023)
178. Lyons, H., Wijenayake, S., Miller, T., Velloso, E.: What’s the appeal? Perceptions of review processes for algorithmic decisions. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–15 (2022)
179. Lyons, H., Miller, T., Velloso, E.: Algorithmic decisions, desire for control, and the preference for human review over algorithmic review. In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 764–774 (2023)
180. Crawford, K., Joler, V.: *Anatomy of an AI system*. Anatomy of an AI System (2018)
181. Widder, D.G., Nafus, D.: Dislocated accountabilities in the “AI supply chain”: modularity and developers’ notions of responsibility. *Big Data Soc.* **10**(1), 20539517231177620 (2023)
182. Cobbe, J., Veale, M., Singh, J.: Understanding accountability in algorithmic supply chains. In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1186–1197 (2023)
183. Balayn, A., Yurrita, M., Rancourt, F., Casati, F., Gadiraju, U.: Unpacking trust dynamics in the LLM supply chain: an empirical exploration to foster trustworthy LLM production & use. In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–20 (2025)
184. Attard-Frost, B., Widder, D.G.: The ethics of ai value chains. *Big Data Soc.* **12**(2), 20539517251340603 (2025)
185. Deng, W.H., Nagireddy, M., Lee, M.S.A., Singh, J., Wu, Z.S., Holstein, K., Zhu, H.: Exploring how machine learning practitioners (try to) use fairness toolkits. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 473–484 (2022)
186. Balayn, A., Yurrita, M., Yang, J., Gadiraju, U., “Fairness toolkits, a checkbox culture?” On the factors that fragment developer practices in handling algorithmic harms. In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 482–495 (2023)
187. Muralikumar, M.D., McDonald, D.W.: Analyzing collaborative challenges and needs of UX practitioners when designing with AI/ML. *Proc. ACM on Hum.-Comput. Interact.* **8**(CSCW2), 1–25 (2024)
188. Zdanowska, S., Taylor, A.S.: A study of UX practitioners roles in designing real-world, enterprise ML systems. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–15 (2022)
189. Liao, Q.V., Subramonyam, H., Wang, J., Wortman Vaughan, J.: Designerly understanding: information needs for model transparency to support design ideation for AI-powered user experience. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–21 (2023)
190. Subramonyam, H., Im, J., Seifert, C., Adar, E.: Solving separation-of-concerns problems in collaborative design of human-AI systems through leaky abstractions. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–21 (2022)
191. Deshpande, A., Sharp, H.: Responsible AI systems: who are the stakeholders? In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 227–236 (2022)
192. Posada, J., Newlands, G., Miceli, M.: Labor, automation, and human-machine communication. *The Sage Handbook of Human-Machine Communication*, pp. 384–391 (2023)
193. Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., Aroyo, L.M.: “Everyone wants to do the model work, not the data work”: data cascades in high-stakes AI. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–15 (2021)
194. Hada, R., Husain, S., Gumma, V., Diddee, H., Yadavalli, A., Seth, A., Kulkarni, N., Gadiraju, U., Vashistha, A., Seshadri, V., et al.: Akal Badi ya Bias: an exploratory study of gender bias in Hindi language technology. In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1926–1939 (2024)

195. Altenried, M.: The platform as factory: crowdwork and the hidden labour behind artificial intelligence. *Capital Class* **44**(2), 145–158 (2020)
196. Balayn, A., Gürses, S.: Beyond debiasing: regulating ai and its inequalities. EDRi report (2021)
197. Jakesch, M., Buşinca, Z., Amershi, S., Olteanu, A.: How different groups prioritize ethical values for responsible AI. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22*, pp. 310–323. Association for Computing Machinery, New York, 2022. <https://doi.org/10.1145/3531146.3533097>
198. Madaio, M.A., Stark, L., Wortman Vaughan, J., Wallach, H.: Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20*, pp. 1–14. Association for Computing Machinery, New York, 2020. <https://doi.org/10.1145/3313831.3376445>
199. Madaio, M., Egede, L., Subramonyam, H., Wortman Vaughan, J., Wallach, H.: Assessing the fairness of AI systems: AI practitioners' processes, challenges, and needs for support. *Proc. ACM Hum.-Comput. Interact.* **6**(CSCW1), 1–26 (2022)
200. Rakova, B., Yang, J., Cramer, H., Chowdhury, R.: Where responsible AI meets reality: practitioner perspectives on enablers for shifting organizational practices. *Proc. ACM Hum.-Comput. Interact.* **5**(CSCW1) (2021). <https://doi.org/10.1145/3449081>
201. Coeckelbergh, M.: Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Sci. Eng. Ethics* **26**(4), 2051–2068 (2020)
202. Balayn, A., Corti, L., Rancourt, F., Casati, F., Gadiraju, U.: Understanding stakeholders' perceptions and needs across the LLM supply chain. Preprint. arXiv:2405.16311 (2024)
203. Deng, W.H., Yildirim, N., Chang, M., Eslami, M., Holstein, K., Madaio, M., Investigating practices and opportunities for cross-functional collaboration around AI fairness in industry practice. In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 705–716 (2023)
204. Balayn, A., Guerses, S.: Misguided: AI regulation needs a shift in focus. *Internet Policy Rev.* **13**(3) (2024)
205. Luitse, D.: Platform power in AI: the evolution of cloud infrastructures in the political economy of artificial intelligence. *Int. Politics Rev.* **13**(2), 1–44 (2024)
206. Van Der Vlist, F., Helmond, A., Ferrari, F.: Big AI: cloud infrastructure dependence and the industrialisation of artificial intelligence. *Big Data Soc.* **11**(1), 20539517241232630 (2024)
207. Narayanan, A., Kapoor, S.: *AI Snake Oil*. Princeton University Press, Princeton (2024)
208. Chien, A.A., Lin, L., Nguyen, H., Rao, V., Sharma, T., Wijayawardana, R.: Reducing the carbon impact of generative AI inference (today and in 2035). In: *Proceedings of the 2nd Workshop on Sustainable Computer Systems*, pp. 1–7 (2023)
209. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models. Preprint. arXiv:2108.07258 (2021)
210. Argote, L., Ingram, P.: Knowledge transfer: a basis for competitive advantage in firms. *Organ. Behav. Hum. Decis. Process.* **82**(1), 150–169 (2000)
211. Carr, N.: *The Glass Cage: How Our Computers Are Changing Us*. WW Norton & Company, New York (2014)
212. Lee, H.-P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., Wilson, N.: The impact of generative AI on critical thinking: self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In: *Proceedings of the 2025 CHI conference on human factors in computing systems*, pp. 1–22 (2025)
213. Erlei, Alexander and Sharma, Abhinav and Gadiraju, Ujwal: Understanding Choice Independence and Error Types in Human-AI Collaboration. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–19 (2024)

Ujwal Gadiraju is a tenured Associate Professor in the Software Technology Department of the EEMCS Faculty, Delft University of Technology. He co-directs the TU Delft "Design@Scale" AI Lab. Ujwal co-leads a research line on human-centred AI and crowd computing at the Web Information Systems group. Previously, he was a postdoctoral researcher at the L3S Research Center in Hannover, Germany, and received a PhD (summa cum laude) in Computer Science from the Leibniz University of Hannover. His work has been recognized with honours, including best paper awards at premier HCI and AI conferences. His current research focuses on advancing human-centred AI through innovative computational techniques and systems that enhance human experiences, align AI systems with human values, promote inclusivity, and foster appropriate human reliance on AI systems.

Agathe Balayn is a Postdoctoral Researcher at Microsoft Research in the FATE team. Prior to that, she obtained her PhD in Computer Science from Delft University of Technology. Her work lies at the intersection of HCI, AI, and critical computing. She has conducted empirical qualitative studies with over 200 AI practitioners and researchers spanning various AI supply chains (e.g. LLMs for customer service activities, computer vision for pig farming). She has surfaced their concerns and practices vis-a-vis (responsible) AI, the factors that shape these practices, and their limitations. Her insights typically result in conceptual frameworks, recommendations for artifacts (e.g. interfaces, toolkits), and governance procedures to guide AI practitioners and bear policy implications.