EJIS
THE EUROPEAN JOURNAL OF INFORMATION SYSTEMS
Kieran Conboy
Virpi Kristiina Tuunainen

European Journal of Information Systems

ISSN: 0960-085X (Print) 1476-9344 (Online) Journal homepage: www.tandfonline.com/journals/tjis20

Complementarity in human-AI collaboration: concept, sources, and evidence

Patrick Hemmer, Max Schemmer, Niklas Kühl, Michael Vössing & Gerhard Satzger

To cite this article: Patrick Hemmer, Max Schemmer, Niklas Kühl, Michael Vössing & Gerhard Satzger (27 Aug 2025): Complementarity in human-AI collaboration: concept, sources, and evidence, European Journal of Information Systems, DOI: 10.1080/0960085X.2025.2475962

To link to this article: https://doi.org/10.1080/0960085X.2025.2475962

EUROPEAN JOURNAL OF INFORMATION SYSTEMS
https://doi.org/10.1080/0960085X.2025.2475962

RESEARCH ARTICLE

Complementarity in human-AI collaboration: concept, sources, and evidence

Patrick Hemmer[a], Max Schemmer[a,b], Niklas Kühl[b,c,d], Michael Vössing[a,b] and Gerhard Satzger[a]

[a]Institute for Information Systems, Karlsruhe Institute of Technology, Karlsruhe, Germany; [b]IBM Germany, Böblingen, Germany; [c]Business & Information Systems Engineering, University of Bayreuth, Bayreuth, Germany; [d]Department of Information Systems, Fraunhofer FIT, Bayreuth, Germany

ABSTRACT

Artificial intelligence (AI) has the potential to significantly enhance human performance across various domains. Ideally, collaboration between humans and AI should result in complementary team performance (CTP)—a level of performance that neither of them can attain individually. So far, however, CTP has rarely been observed, suggesting an insufficient understanding of the principle and the application of complementarity. Therefore, we develop a general concept of complementarity and formalize its theoretical potential as well as the actual realized effect in decision-making situations. Moreover, we identify information and capability asymmetry as the two key sources of complementarity. Finally, we illustrate the impact of each source on complementarity potential and effect in two empirical studies. Our work provides researchers with a comprehensive theoretical foundation of human-AI complementarity in decision-making and demonstrates that leveraging these sources constitutes a viable pathway towards designing effective human-AI collaboration, i.e., the realization of CTP.

ARTICLE HISTORY
Received 11 July 2023
Accepted 28 February 2025

KEYWORDS
Human-AI complementarity; human-AI decision-making; human-AI teams; human-AI collaboration; complementarity potential; complementarity effect

1. Introduction

The increasing capabilities of artificial intelligence (AI) have paved the way for collaborating with humans and supporting them in a wide range of domains. Examples include decision support for humans in application areas such as customer services (Vassilakopoulou et al., 2023), medicine (Jussupow et al., 2021), law (Mallari et al., 2020), finance (Day et al., 2018), and industry (Stauder & Kühl, 2022). With AI decisions becoming increasingly accurate, there is an obvious temptation to fully rely on them and to automate decision tasks. However, this approach often falls short of realizing even better performance by combining and integrating the unique strengths of the individual members in a human-AI team (Seeber et al., 2020). The recent emergence of large language models illustrates this (Vaccaro et al., 2024): While applications like ChatGPT often provide helpful, but not always correct results, a human decision-maker can collaborate with the system to, for example, override erroneous responses in order to achieve superior task performance (Vaccaro et al., 2024). Similarly, in the medical domain, both AI models and physicians are able to produce diagnoses individually. It has, however, been demonstrated that humans and AI models could make different errors (Geirhos et al., 2021; Steyvers et al., 2022) so that they may realize superior results when "teamed up": For instance, the AI model might detect patterns in large amounts of data that humans might not discover easily, while humans might excel at the causal interpretation and intuition required to understand these patterns (Lake et al., 2017).

This complementarity as the "quality of being different but useful when combined" (Cambridge Dictionary, 2024) has inspired researchers to investigate how humans' and AI's individual abilities could be leveraged to achieve superior team performance compared to either one performing the decision task independently. Such an outcome is defined as *complementary team performance (CTP)* (Bansal et al., 2021). This phenomenon is of increasing interest as recent studies tell only part of the story: On the one hand, various studies have demonstrated that human-AI teams are able to outperform human individuals (Alufaisan et al., 2021; Inkpen et al., 2023; Liu et al., 2021; Sarkar et al., 2023)—often not analyzing, though, whether they also surpass the AI model's individual performance (Bansal et al., 2021). On the other hand, while many settings exist in which humans still show better task performance (Grace et al., 2018, 2024), recent studies also demonstrate evidence for human-AI teams outperforming AI models (Dvijotham et al., 2023; Fügener et al., 2022; Ma et al., 2023).

To make things even more complex, the particular design of human-AI collaboration in decision-making affects humans and their contributions within the

CONTACT Niklas Kühl kuehl@uni-bayreuth.de

Supplemental data for this article can be accessed online at https://doi.org/10.1080/0960085X.2025.2475962.

team, e.g., the use of unique human knowledge (Fügener et al., 2021), the self-assessment of human capabilities (Fügener et al., 2022), the adjustments of mental models (Bauer et al., 2023), the incentive for "active rethinking" (Lu & Zhang, 2024), or human task performance and learning (Förster et al., 2024). In summary, current research still misses compelling explanations for the success of human-AI teams as well as a systematic understanding of complementarity when the team performance is measured against the performance of *both* team members individually. Thus, concentrating on decision-making as an important application of human-AI collaboration (Lai et al., 2023), we pursue the following two research questions in this work:

RQ1: How can we model human-AI collaboration in decision-making to enable a more nuanced understanding of the synergetic potential in a human-AI team?

RQ2: What factors contribute to complementary team performance?

We address these research questions by developing a conceptualization of human-AI complementarity that introduces and quantifies *complementarity potential (CP)* and *complementarity effect (CE)* and outlines the two key sources of complementarity—information and capability asymmetry. In detail, we argue that both complementarity potential and complementarity effect consist of an inherent and a collaborative component. Whereas the first component captures decision-making synergies that, for each task instance, can be attributed to the individually more accurate team member within the human-AI team, the second component captures decision-making synergies that only emerge through collaboration resulting in team decisions which are more accurate than each of the individual ones. We demonstrate the application and the value of our conceptualization in two experimental studies—leveraging the two sources of complementarity, i.e., information and capability asymmetry within the human-AI team. In both studies, humans collaborate with an AI model to conduct decision-making tasks. The AI model provides independent decision suggestions that humans can incorporate into their judgment to derive a final team decision. In the first study, we choose the domain of real estate valuation to investigate information asymmetry. We train an AI model to predict real estate prices based on tabular data. Humans receive its suggestions and also have access to a photograph of the real estate. They can use both to arrive at a final team decision. In the second study, we select the context of image classification to analyze the impact of capability asymmetry between humans and AI. We train two AI models whose capability gaps differ from those of the human decision-maker. In both studies, we apply our conceptualization to demonstrate that both of these sources increase the inherent component of complementarity potential as well as the realized effect, resulting in CTP.

To summarize, we make the following contributions to the state of knowledge in information systems (IS) research: First, we conceptualize human-AI complementarity as a means to comprehensively analyze and design human-AI collaboration in decision-making. Second, we scrutinize information and capability asymmetries as sources of complementarity. Third, we demonstrate the value and application of our concepts and the sources' potential impact in two behavioral experiments. This should provide IS researchers with a better understanding and methodological support when purposefully designing human-AI collaboration in decision-making for more effective outcomes—thereby supporting the development of hybrid intelligence (Dellermann et al., 2019) and advocating the "AI *with* human" (opposed to "AI *vs.* human") perspective in societal debates on the future of work (Huysman, 2020).

In the remainder of this work, we first outline the relevant background and related work in Section 2. In Section 3, we derive the conceptualization of human-AI complementarity. In Section 4, we empirically illustrate our conceptualization's utility in two experimental studies. We discuss our results and conclude the work in Section 5.

2. Theoretical foundations and related work

In this section, we elaborate on the key concepts and on existing work that provide the foundation for a formal conceptualization of complementarity. We first cover existing differences between humans and AI models[1] as sources of complementarity, then review existing literature on human-AI collaboration in decision-making as the means to harness complementarity, and finally summarize the current state of knowledge on human-AI complementarity.

2.1. Sources of complementarity: information and capability asymmetry

In the following, we assume that both human and AI can independently produce solutions to a particular decision problem. Collaboration would not be of interest if both always generated identical decisions. However, human and AI typically make *different* types of errors (Geirhos et al., 2021; Steyvers et al., 2022)—a phenomenon that very generally can be traced back to two key sources. To illustrate this, we employ a simple Input-Processing-Output Model for decision-making (Figure 1)—as used both in IS (D'Arcy

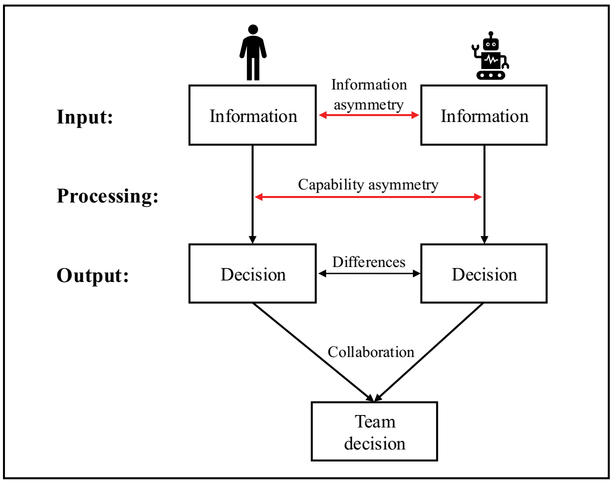


Figure 1. Asymmetries in decision-making between human and AI as sources of complementarity.

et al., 2014; Parsons & Wand, 2012) and social psychology (Gladstein, 1984; Hackman & Morris, 1975): Any team player (human or AI) draws on and processes certain sets of information to derive a decision. Discrepancies in the decision outcome may either stem from different levels of available information (*information asymmetry*) or from different capabilities to process this information (*capability asymmetry*). The two asymmetries may then be exploited for a (hopefully) improved team decision resulting in CTP.

2.1.1. Information asymmetry

Often, humans and AI have access to different sets of information as decision input: On the one hand, AI models are trained on a well-defined, limited, and digitally available set of data (LeCun et al., 2015). Humans, however, may also have access to information that—due to technical or economic reasons—is not digitized and, thus, not usable for an AI model (Ibrahim et al., 2021). They may use contextual information or information on rare events for a more holistic decision-making setting. For example, AI models in the medical domain derive diagnoses from X-ray scans by analyzing the pixels of the image (Irvin et al., 2019). Human radiologists, however, may also draw on information from direct interaction with patients or from access to medical records. On the other hand, the AI may also have access to information not available to the human decision-maker: A driving assistance system may base its actions on sensor data, e.g., from lidar systems that are not accessible by the human while driving (Li & Ibanez-Guzman, 2020).

2.1.2. Capability asymmetry

Even if the information would be identical, different outcomes may be driven by different internal modes of information processing between human and AI (Kühl, Goutier, et al., 2022). Such capability asymmetries emerge as AI models encode relationships inferred from training data, whereas humans employ compositional mental models that encode beliefs about the physical and social world (Lake et al., 2017; Rastogi et al., 2023). While humans are capable of conducting decision-making tasks already after a few trials, AI models require vast amounts of training data to become capable of accurately performing decision-making tasks (Gopnik & Wellman, 2012; Kühl, Goutier, et al., 2022; Lake et al., 2017; Tenenbaum et al., 2011). Similarly, humans have often gained experiences with regard to a particular decision-making task continuously over their lifetime, while AI models' experiences are limited to the task instances seen during model training and may only be updated from time to time (Dellermann et al., 2019; Rastogi et al., 2023). On the other hand, AI may excel in digesting vast amounts of information much faster than humans could, and its processing exhibits a greater capability to perceive even small variations in data (Findling & Wyart, 2021).

Overall, both asymmetries may give rise to performance synergies when collaborating in human-AI teams. A traditional example is forecasting theory (Sanders & Ritzman, 1991, 1995, 2001): Sanders & Ritzman (1995) analyze the effects of combining statistical forecasts with predictions of a human with access to contextual information and find that such combinations could positively impact forecast accuracy. Composing suitable teams to optimize team performance has been investigated for human-only teams (Horwitz, 2005), but lately has also been applied to human-AI teams (Hemmer et al., 2022). While research has shown that such asymmetries can improve team performance (Simons et al., 1999; Q. Zhang et al., 2022), they also bear the risk of negative effects, i.e., performance degradation (Dougherty, 1992): Heterogeneous and interdisciplinary human-only teams may enjoy benefits from higher levels of individual problem-solving creativity that may be outweighed, though, by the difficulties to effectively communicate with each other (Ancona & Caldwell, 1992).

2.2. Means for harnessing complementarity: human-AI collaboration

Many terms are used to describe the interplay between humans and AI. Common ones are *human-AI team* (Seeber et al., 2020), *human-AI collaboration* (Vössing et al., 2022), and *human-AI decision-making* (Lai et al., 2023). These are interrelated concepts that emphasize combining the complementary qualities of humans and AI. The notion of a *human-AI team* refers to an organizational setup in which AI systems are increasingly considered equitable team members rather than support tools for humans—since AI can perform a continuously growing number of tasks independently (Endsley, 2023; Seeber et al., 2020). *Human-AI collaboration* is the process in which these teams work together in a synergistic manner to achieve shared goals, e.g., with the AI providing recommendations or insights, and humans guiding and refining the AI-generated outputs (Terveen, 1995; Vössing et al., 2022). *Human-AI decision-making*, which is the focus of this work, refers specifically to human-AI collaboration in decision-making tasks (Lai et al., 2023).

The increasing abilities of AI have contributed to its use in a growing number of application domains (Kleinberg et al., 2018; McKinney et al., 2020; Mikalef & Gupta, 2021; Vassilakopoulou et al., 2023). Consequently, AI-based technologies are employed in processes and systems with varying degrees of human involvement, ranging from autonomous decision-making (Rinta-Kahila et al., 2022) to just auxiliary support for the ultimate human decision-maker (Bansal et al., 2021; Buçinca et al., 2020; Lai et al., 2020; Liu et al., 2021). In this context, unintended or unfair outcomes, e.g., AI-based systems' decisions that benefit certain individuals more than others (Kordzadeh & Ghasemaghaei, 2022), have ignited a debate on the degree of autonomy granted to AI to ensure responsible outcomes (Mikalef et al., 2022). To alleviate possible detrimental effects, configurations have been suggested that keep humans in the decision-making loop (Grønsund & Aanestad, 2020; Herath Pathirannehelage et al., 2024; Mikalef et al., 2022).

Researchers have been devoting significant efforts to better understand human-AI decision-making and to design the collaboration in a way that ultimately achieves CTP (Hemmer et al., 2021; Lai et al., 2023). Overall, a wide range of behavioral experiments seek to help us understand how humans make decisions within human-AI teams (Alufaisan et al., 2021; Bansal et al., 2021; Buçinca et al., 2020; Carton et al., 2020; Fügener et al., 2021, 2022; Lai et al., 2020; Liu et al., 2021; Vaccaro et al., 2024; Reverberi et al., 2022; van der Waa et al., 2021; Q. Zhang et al., 2022). A key emerging concept in this space is that of human reliance on AI advice that needs to be appropriately calibrated to ensure effective decision-making (Buçinca et al., 2020; He et al., 2023; Kunkel et al., 2019; Schemmer et al., 2022; Schoeffer et al., 2023; Yu et al., 2019; Y. Zhang et al., 2020). To assist the human in judging the AI's decision quality, the AI can provide information about the decision's uncertainty (Fügener et al., 2021; Y. Zhang et al., 2020) or deliver various types of explanations that shed light on its decision-making rationale (Adadi & Berrada, 2018; Bauer et al., 2023).

A closer look at quantitative studies on human-AI decision-making reveals that, in general, human performance increases when supported by high-performing AI models. In the vast majority of cases, however, the team performance remains inferior to that of the AI model when performing the task alone (Hemmer et al., 2021; Vaccaro et al., 2024). This means that joint decision-making currently does not lead to the realization of the full complementarity potential: Humans often do not show appropriate reliance by contributing their own decision capabilities in the right places. While recent studies have shown that the performance of human-AI teams *can* improve beyond that of the individual team members (Dvijotham et al., 2023; Fügener et al., 2022; Ma et al., 2023), the underlying mechanisms *why* performance synergies often fail to materialize are still poorly understood. Research has explored different paths: First, humans' ability to exert appropriate reliance depends on the overall decision-making situation, e.g., whether it is possible to ex post verify the correctness of the decisions (de Véricourt & Gurkan, 2023). Second, "imperfections" on the human side may contribute to this: Humans can struggle to correctly assess their own capabilities in comparison to that of the AI (Fügener et al., 2022). They may develop implicit biases against the AI that inhibit their willingness to rely on its advice (Turel & Kalhan, 2023), or they may start to mirror the AI's behavior by following its advice even when it is incorrect (Fügener et al., 2021). Third, humans may be misled by signals from the AI that were originally intended to help them better assess its decision quality. Additional explanations may not be correct (Morrison et al., 2024), or they may distort humans' situational balancing of available information, leading to misconceptions and suboptimal decisions (Bauer et al., 2023). Overall, further research is required to provide the means for developing a holistic understanding of the synergetic potential between humans and AI—an objective this work pursues by proposing a conceptualization of human-AI complementarity.

2.3. Concepts related to human-AI complementarity

Complementarity between humans and AI is discussed as part of several closely related paradigms:

Intelligence augmentation, *human-machine symbiosis*, and *hybrid intelligence.*

Intelligence augmentation is defined as "enhancing and elevating human's ability, intelligence, and performance with the help of information technology" (Zhou et al., 2021, p. 245). It follows the idea that machines use their abilities to assist humans, not necessarily to achieve CTP, but to improve human objectives. *Human-machine symbiosis* is a paradigm that envisions deepening the collaborative connection between humans and AI. It is based on the notion of a symbiotic relationship between both and considers them as a common system rather than two separate entities with the aim of becoming more effective together than working separately (Licklider, 1960). It also makes the assumption that both entities can offer different capabilities that can be leveraged to overcome human restrictions and to reduce the time needed to solve problems (Gerber et al., 2020; Jain et al., 2021). *Hybrid intelligence* pursues the idea of combining human and AI team members in the form of a socio-technical ensemble. We refer to the work of Dellermann et al. (2019, p. 640), who define hybrid intelligence as "the ability to achieve complex goals by combining human and artificial intelligence, thereby reaching superior results to those each of them could have accomplished separately, and continuously improve by learning from each other".

Nevertheless, existing studies under these labels do not provide theoretical views of human-AI complementarity. Articles by Donahue et al. (2022), Steyvers et al. (2022), and Rastogi et al. (2023) are the only works to theorize about human-AI complementarity. Donahue et al. (2022) discuss scenarios in which CTP could occur by considering fairness aspects. Steyvers et al. (2022) derive a framework for combining individual decisions and different types of confidence scores from humans and AI models. Lastly, Rastogi et al. (2023) propose a taxonomy of human and AI strengths together with the notion of across- and within-instance complementarity. All three differ from our approach as they do not distinguish between complementarity potential (*CP*) and complementarity effect (*CE*) with their respective components. Moreover, they do not empirically analyze human-AI decision-making in behavioral experiments.

3. Conceptualization of human-AI complementarity

In this section, we first introduce the fundamental notion of human-AI complementarity and formalize our decision-making situation as a basis for further analysis. Subsequently, we introduce and formalize the concepts of complementarity potential and effect, and relate them to the underlying sources of complementarity.

3.1. The principle of human-AI complementarity

We first motivate the underlying idea of complementarity which drives effective human-AI decision-making. In this work, we focus on the performance on decision-making tasks that humans and AI can conduct independently—recognizing that AI models have elevated above pure decision support for humans. However, since neither humans nor AI are perfect, discrepancies in access to information or in capabilities could be leveraged to generate superior outcomes in a human-AI team. Figure 2 illustrates the situation in a simple example for a set of decisions: The decision-making task comprises 25 instances that each have a set of possible discrete outcomes only one of which is correct. The number of incorrect decisions measures the performance of each individual team member and the human-AI team, respectively. The AI makes 13 incorrect decisions, while the human errs at 15 task instances when conducting the task independently. 5 of the instances can neither be solved by the AI nor the human on their own. Consequently, if the human-AI team were just to pick the correct decision of either the AI or the human, the team could correctly solve 20 instances and would only miss the 5 that none of them can solve. In other words, relying on the correct individual decisions of *each* team member improves the result compared to the AI acting alone (as the individually better performing team member): While the AI still shows 13 incorrect decisions when acting independently, teaming up with the human can reduce this number to 5—realizing an improvement potential of 8 decisions. Moreover, it is also conceivable that—while none of the team members is correct

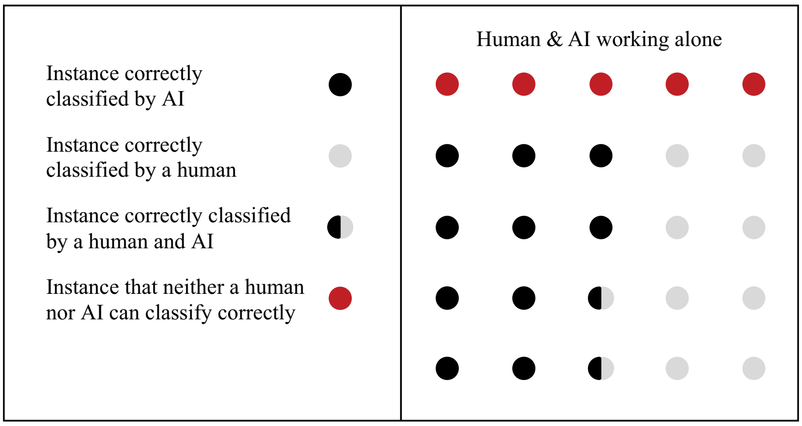

Instance correctly classified by AI
Instance correctly classified by a human
Instance correctly classified by a human and AI
Instance that neither a human nor AI can classify correctly
Human & AI working alone

Figure 2. Illustration of the principle of human-AI complementarity based on different (in)correct decisions that human and AI can make.

individually—the interaction between the team members may allow the generation of correct team decisions even for the remaining 5 red task instances in Figure 2 representing an additional improvement potential:[2] Team members may recognize through collaboration that they have different information or capabilities that they can jointly apply. Let us, e.g., assume that a task would be to deliver a diagnosis on cancer. Both a human radiologist and an AI analyzing patient data might each individually render a wrong diagnosis, e.g., "malign cancer" conjectured by the human and "no cancer" proposed by the AI. If, however, the human might learn about the AI's rationale (e.g., regions in the X-ray scan crucial for the decision) or the AI might learn about "side" information that the radiologist has on the patient history, they jointly may arrive at the correct diagnosis "benign cancer".

For human-AI teams to achieve CTP, it is essential that they manage to realize these improvement potentials. If their information and capabilities could be adequately combined, the team performance in such situations would be superior to their individual performances (Rastogi et al., 2023).

In the introductory example, performance is captured by the absolute number of errors. Depending on the application context, more intricate measures could also be applied, e.g., precision or recall in classification tasks (e.g., when analyzing radiology images in health care), mean absolute error in prediction tasks (e.g., when making sales forecasts for inventory management), or more complex compound metrics (e.g., when weighting multiple dimensions of interest).

In reality, the performance of humans and AI varies depending on the task and the application domain. Enabled by advances in AI research over the last few years, there has been an increasing number of tasks where the performance of AI has reached or even exceeded that of humans (Afshar et al., 2022; Bubeck et al., 2023; Silver et al., 2018). However, many applications remain in which human performance remains superior (Brynjolfsson et al., 2018; Grace et al., 2018, 2024). For the motivating example, we chose a setting in which the AI makes fewer errors—as a reflection of the developments in AI model performance over the last years. However, the conceptualization that we develop in this section is independent of the performance relationship between human and AI.

3.2. Human-AI decision-making setting

Let us first define the human-AI decision-making setting illustrated in Figure 2, which is the foundation of this work: A decision task $T = \left\{\left(x^{(i)}, y^{(i)}\right)\right\}_i^N$ is a set of N instances $x^{(i)} \in X$ with corresponding ground truth labels $y^{(i)} \in Y$ denoting the correct results. In this section, we use the term decision for both classification and prediction tasks. The ground truth, i.e., the correct decision, might not be known at the time of the decision, but can be determined and revealed later. Both a human decision-maker H and a machine learning model, which we denote as AI, are capable of independently producing a decision for each task instance. For any given instance $x^{(i)}$, the human and the AI will independently derive decisions $\hat{y}_H^{(i)}$ and $\hat{y}_{AI}^{(i)}$. In a scenario where the human and the AI might collaborate in respect of their decision-making, a collaboration mechanism $I\left(\hat{y}_H^{(i)}, \hat{y}_{AI}^{(i)}\right)$ integrates their decisions into a final team decision $\hat{y}_I^{(i)}$. We note that this decision might be different from each individual decision.

Each decision's quality is measured by its deviation ("loss") from the ground truth—by a loss function l bounded in R^+. This function serves as a generic measure of performance that could take different forms with different decision problems, e.g., an error rate used in classification tasks (like the number of wrong or unsolved task instances in Figure 2). For any instance $x^{(i)}$, losses $l_D^{(i)}$ with $D \in \{H, AI, I\}$ are given by $l_H^{(i)}$ for the human decision, $l_{AI}^{(i)}$ for the AI decision, and $l_I^{(i)}$ for the integrated team decision. This results in overall losses L_D for the entire task by averaging all the available instances:

$$L_D = \frac{1}{N}\sum_{i=1}^{N} l_D^{(i)}\left(\hat{y}_D^{(i)}, y^{(i)}\right) \; with \; D \in \{H, AI, I\}. \quad (1)$$

3.3. Complementarity potential

From a decision-theoretic perspective, the vision of human-AI decision-making is to attain a superior team performance compared to the human and the AI conducting the task individually—providing the fundamental reason for forming human-AI teams (Rastogi et al., 2023). In our context, the human-AI team reaches *complementary team performance (CTP)* when the loss of the team is strictly smaller than that of the human and the AI individually (Bansal et al., 2021):

$$CTP = \begin{cases} 1, & L_I < \min(L_H, L_{AI}), \\ 0, & otherwise. \end{cases} \quad (2)$$

In addition to a binary task outcome, we propose the notion of *complementarity potential* (*CP*) to measure the discrepancy between the overall loss of the individually better performing team member $T^* \in \{H, AI\}$ with $L_{T^*} = \min(L_H, L_{AI})$ and perfect decisions for all instances of a task, i.e., the selection of the ground truth, with an overall loss of 0:

$$CP = L_{T^*} = \min(L_H, L_{AI}). \quad (3)$$

In our introductory example in Figure 2, the overall loss is quantified by the number of wrong decisions. Consequently, the complementarity potential amounts to 13, which is given as the minimum number of individual errors (13 for the AI, 15 for the human).[3] It should be noted that only 8 task instances of this potential can be realized by picking the better individual decision, while 5 task instances cannot be solved individually, but only—if at all—in collaboration, resulting in a decision that neither AI nor human could come up with on their own. We reflect this by distinguishing two components: *Inherent* and *collaborative* complementarity potential.

Inherent complementarity potential represents improvement potential, i.e., loss reductions, that—from the perspective of the overall more accurate team member T^*—could be contributed by any superior decisions on the instance level by the overall less accurate team member. This means that the team decision is confined to the solutions contributed by one of the team members—in the introductory example (Figure 2) represented by the 20 "gray and black" task instances solvable by at least human or AI:

$$\hat{y}_I^{(i)} = I\left(\hat{y}_H^{(i)}, \hat{y}_{AI}^{(i)}\right) \in \left\{\hat{y}_H^{(i)}, \hat{y}_{AI}^{(i)}\right\}. \tag{4}$$

Collaborative complementarity potential, on the other hand, signifies improvement potential, i.e., loss reductions, that go beyond the individual team solutions by generating "new" knowledge. It may be noted that this is only possible if there is interaction between the team members enabling them to learn from the result of the partner and realize integrated values $\hat{y}_I^{(i)} \in Y$ *different* from the individual ones ($\hat{y}_H^{(i)}$ and $\hat{y}_{AI}^{(i)}$). In the introductory example (Figure 2), this is captured by the 5 "red" task instances not solvable by either team member alone.

We illustrate both potentials from a general perspective on a continuous range by looking at the individual losses of human and AI for a single task instance in Figure 3. Without loss of generality, we assume that the AI is the overall better performing individual team member ($L_{AI} \leq L_H$). If for this particular instance the overall inferior team member, i.e., the human, could help reduce the loss with his/her decision, we denote *inherent* complementarity potential (scenario 1). The remaining loss is unavoidable if—according to Equation (4)—the team decision is restricted to one of the team members' individual decisions. Thus, the loss of the better performing team member for a task instance constitutes the *collaborative* complementarity potential. This potential could only be exploited if the team members' collaboration yields new insights for this task instance that were not available for the individual decisions before the collaboration (scenario 1 and 2), resulting in a team decision that incurs a loss lower than that of each team member (human and AI) individually.

Formally, inherent complementarity potential CP^{inh} can be calculated by aggregating all potential loss improvements on the instance level where the overall worse performing team member can help improve the team result:

$$CP^{inh} = \frac{1}{N}\sum_{i=1}^{N} \begin{cases} \max(0, l_{AI}^{(i)} - l_H^{(i)}), & L_{AI} \leq L_H, \\ \max(0, l_H^{(i)} - l_{AI}^{(i)}), & L_{AI} > L_H. \end{cases} \tag{5}$$

The collaborative complementarity potential CP^{coll} can be calculated by aggregating the remaining minimum losses per task instance that the team members incur individually:

$$CP^{coll} = \frac{1}{N}\sum_{i=1}^{N} \min\left(l_H^{(i)}, l_{AI}^{(i)}\right). \tag{6}$$

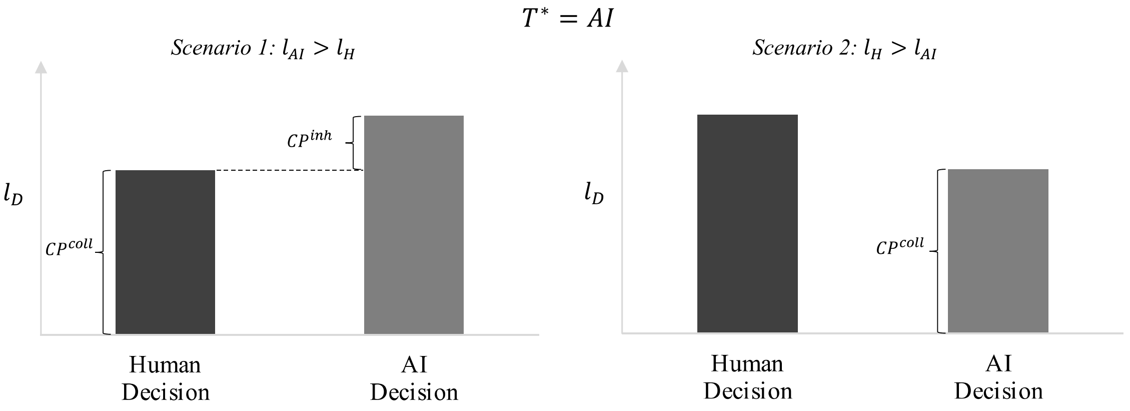


Figure 3. Complementarity potential (*CP*) split into inherent (CP^{inh}) and collaborative (CP^{coll}) components for a single instance with better human performance (left) and better AI performance (right) in respect of $T^* = AI$. l_D denotes the instance-specific loss with $D \in \{H, AI\}$ with a lower loss indicating better performance for the same instance.

The inherent and collaborative components are additive and together form the total complementarity potential CP (see Appendix A for additional details):

$$CP = CP^{inh} + CP^{coll}. \tag{7}$$

In our introductory example, the total complementarity potential of 13 instances can be differentiated into an inherent complementarity potential (CP^{inh}) of 8 instances (for which the human team member can contribute the correct solution), and a collaborative complementarity potential (CP^{coll}) of 5, with none of the team members arriving at the correct decision individually.

3.4. Complementarity effect

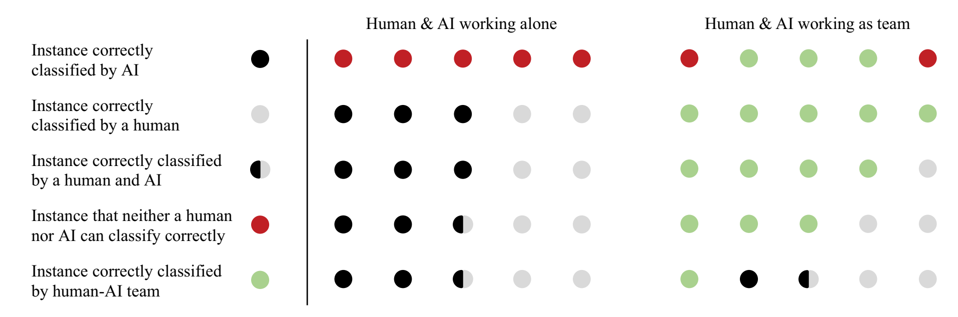
In real-world collaboration scenarios between a human and AI, it is, of course, unlikely that the entire complementarity potential will be exploited. We, therefore, introduce the *complementarity effect (CE)* as that part of this potential that is actually realized by the integrated team decision. Measuring and dissecting this effect will allow observed human-AI decision-making settings to be analyzed in greater detail, in order to infer conclusions about the collaboration's effectiveness, and to purposefully develop and compare collaboration designs and mechanisms. Analogous to the complementarity potential in Equation (3), the realized complementarity effect accounts for the difference between the average loss of the overall individually better team member and that of the integrated team decision ($L_I > 0$):

$$CE = \min(L_H, L_{AI}) - L_I. \tag{8}$$

We expand our introductory example (Figure 2) in Figure 4 by incorporating (hypothetical) integrated human-AI team decisions for all instances. Let us assume that the human-AI team makes 9 incorrect decisions, compared to 13 errors by the AI and 15 errors by the human when acting independently. Thus, this collaboration generates a complementarity effect of 4—realizing 31% of the full complementarity potential of 13.

The complementarity effect measures loss improvements—from the perspective of the overall more accurate team member T^*—that are realized by the integration of individual decisions into a team decision. Analogous to inherent or collaborative complementarity potential, we can split the complementarity effect into the same two categories. Figure 5 illustrates the different scenarios that could materialize for a particular task instance in terms of the losses caused by the solutions of human, AI, and the (integrated) team for the same task instance. Again, we assume, without loss of generality, the AI to be the overall better performing individual team member ($L_{AI} \leq L_H$). If there is *inherent* complementarity potential (i.e., the overall inferior human is more knowledgeable for the particular task instance), this potential may be realized either partially ($l_{AI} > l_I > l_H$, scenario 1), fully ($l_{AI} > l_H > l_I$, scenario 2), or not at all ($l_I > l_{AI}$, scenario 3). No inherent complementarity potential exists where the overall better performing team member also dominates in the particular task instance (scenario 4).

In addition, the collaborative complementarity potential (as the smaller loss of the team members in each scenario) could be tapped into. While in scenario 1 only a fraction of the inherent complementarity potential is realized, scenario 2 not only fully exploits the inherent complementarity potential, but also taps into some of the collaborative complementarity potential. Neither scenarios 3 nor scenario 4 realize any inherent complementarity potential, but solely contribute collaborative complementarity effects: In scenario 3 the integrated solution performs even worse than that of the inferior team member—resulting in a *negative* collaborative complementarity effect, i.e.,


Human & AI working alone
Human & AI working as team
Instance correctly classified by AI
Instance correctly classified by a human
Instance correctly classified by a human and AI
Instance that neither a human nor AI can classify correctly
Instance correctly classified by human-AI team

	Inherent	Collaborative	Total
Theoretical: Complementarity Potential	8	5	13
Realized: Complementarity Effect	3	1	4

Figure 4. Illustration of (theoretical) complementarity potential and (realized) complementarity effect for a hypothetical situation extending the introductory example in Figure 2.

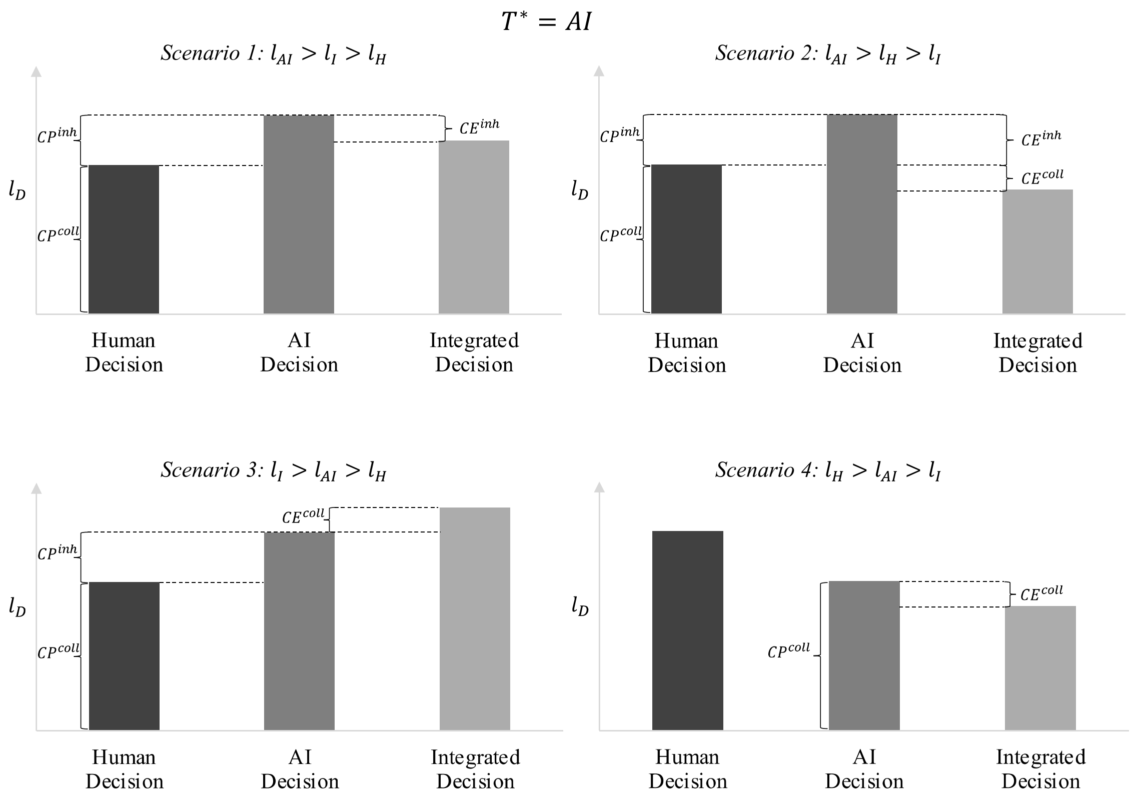


Figure 5. Illustration of (theoretical) complementarity potential (CP) and (realized) complementarity effect (CE) in respect of different loss scenarios for a single instance—assuming, without loss of generality, that the AI performs better overall ($T^* = AI$). l_D denotes the instance-specific loss with $D \in \{H, AI, I\}$—with a lower loss indicating better performance for the same instance.

the collaboration actually worsens the outcome. Conversely, in scenario 4, the integrated solution outperforms the better team member, generating a *positive* collaborative complementarity effect.

In general, we can aggregate the complementarity effects across all instances of a task and summarize both cases (AI or human with overall better performance) and the scenarios above (depending on the instance performance of human, AI, and human-AI team):

$$CE^{inh} = \frac{1}{N}\sum_{i=1}^{N}\begin{cases} l_{AI}^{(i)} - l_{I}^{(i)}, & L_{AI} \le L_H \ and \ l_{AI}^{(i)} > l_{I}^{(i)} \ge l_{H}^{(i)}, \\ l_{AI}^{(i)} - l_{H}^{(i)}, & L_{AI} \le L_H \ and \ l_{AI}^{(i)} > l_{H}^{(i)} > l_{I}^{(i)}, \\ l_{H}^{(i)} - l_{I}^{(i)}, & L_H < L_{AI} \ and \ l_{H}^{(i)} > l_{I}^{(i)} \ge l_{AI}^{(i)}, \\ l_{H}^{(i)} - l_{AI}^{(i)}, & L_H < L_{AI} \ and \ l_{H}^{(i)} > l_{AI}^{(i)} > l_{I}^{(i)}, \\ 0, & otherwise. \end{cases} \quad (9)$$

$$CE^{coll} = \frac{1}{N}\sum_{i=1}^{N}\begin{cases} l_{H}^{(i)} - l_{I}^{(i)}, & l_{AI}^{(i)} \ge l_{H}^{(i)} > l_{I}^{(i)}, \\ l_{AI}^{(i)} - l_{I}^{(i)}, & l_{H}^{(i)} > l_{AI}^{(i)} > l_{I}^{(i)}, \\ l_{AI}^{(i)} - l_{I}^{(i)}, & L_{AI} \le L_H \ and \ l_{I}^{(i)} > l_{AI}^{(i)}, \\ l_{H}^{(i)} - l_{I}^{(i)}, & L_H < L_{AI} \ and \ l_{I}^{(i)} > l_{H}^{(i)}, \\ 0, & otherwise. \end{cases} \quad (10)$$

Analogous to Equation (7), the inherent and collaborative components add up to the total complementarity effect (see Appendix A for additional details):

$$CE = CE^{inh} + CE^{coll}. \quad (11)$$

In our extended introductory example in Figure 4, we find that of the 8 task instances offering inherent complementarity potential, 3 could be realized as an inherent complementarity effect (CE^{inh}) by taking the human's individual decision suggestions into account. Of the 5 task instances for which neither the human nor the AI could individually make a correct decision, collaboration enabled the human-AI team to make correct decisions regarding 3 task instances. However, also 2 task instances that the AI alone could have performed correctly are subject to an erroneous decision due to the collaboration. Consequently, the collaborative complementarity effect (CE^{coll}) amounts to 1 and the total complementarity effect (CE) to 4.

3.5. Sources of complementarity: the impact on complementarity potential and effect

In the motivating example, the human and the AI make erroneous decisions for different task instances allowing for possible performance synergies that may be attainable through collaboration. We used this example to introduce the measures of complementarity potential and effect.

Differences in human and AI decision-making can arise from information and capability asymmetries as introduced earlier (Section 2.1). These asymmetries are fundamental for the existence of complementarity. They cause different decisions to be taken by humans and AI on an instance level and, therefore, influence the complementarity potential and the realized complementarity effect including both components. Leveraging information asymmetry means making information available to the team: Advantages can be captured as inherent (when the team decision for a task instance corresponds to the individual decision of the team member with "better" information) or as collaborative complementarity potential/effect (when the information of both team members is joined resulting in new insights and a team decision different from the individual ones). Similarly, capability asymmetries may contribute to inherent (when the team decision corresponds to the individual decision of the team member that is more capable for the particular task instance) or collaborative complementarity potential/effect (when capabilities complement each other, e.g., human experience and AI computational power, resulting in a team decision that differs from the individual ones).

It may be noted, though, that realizing complementarity potential (and, thus, achieving CTP) involves managing a number of trade-offs: Information and capabilities are typically not distributed in a way that either the human or the AI dominates across all existing task instances (Geirhos et al., 2020, 2021; Kühl, Goutier, et al., 2022). There may be pieces of information that only the human or only the AI has access to: In our earlier radiology example, this may be contextual information about a patient that the physician has and a large variety of cases in the training set that only the AI is able to access. In addition, also information may be traded off against capabilities: Assuming the radiologist has dominating information, it may be outweighed by the capabilities of the AI to automatically digest and evaluate the information available to it in real time.

We summarize the notion of complementarity, its sources, and the measures of complementarity potential and effect in the conceptual framework in Figure 6. Each source can affect both the "inherent" and "collaborative" components of complementarity potential and effect. Our conceptualization presented in this section is intended to provide a deeper understanding of the phenomenon as well as to provide a concrete measurement construct for systematically harnessing complementarity.

To illustrate and evaluate this framework, we now design two behavioral experiments in which humans and AI make decisions on their own and jointly for a set of task instances. We investigate how the presence of information (Section 4.1) and capability (Section 4.2) asymmetry affect the team's complementarity potential, the realized complementarity effect as well as the joint team performance.

4. Experimental studies

To demonstrate the proposed conceptualization's value and application, and to further investigate human-AI decision-making in the presence of the identified sources of complementarity—information and capability asymmetry—we conducted two behavioral experiments. Specifically, we focused on a team setting in which a human decision-maker has access to AI advice and is subsequently responsible for making the final team decision based on his/her judgment and that of the AI (Green & Chen, 2019). In this collaboration setup, the team decision either matches the

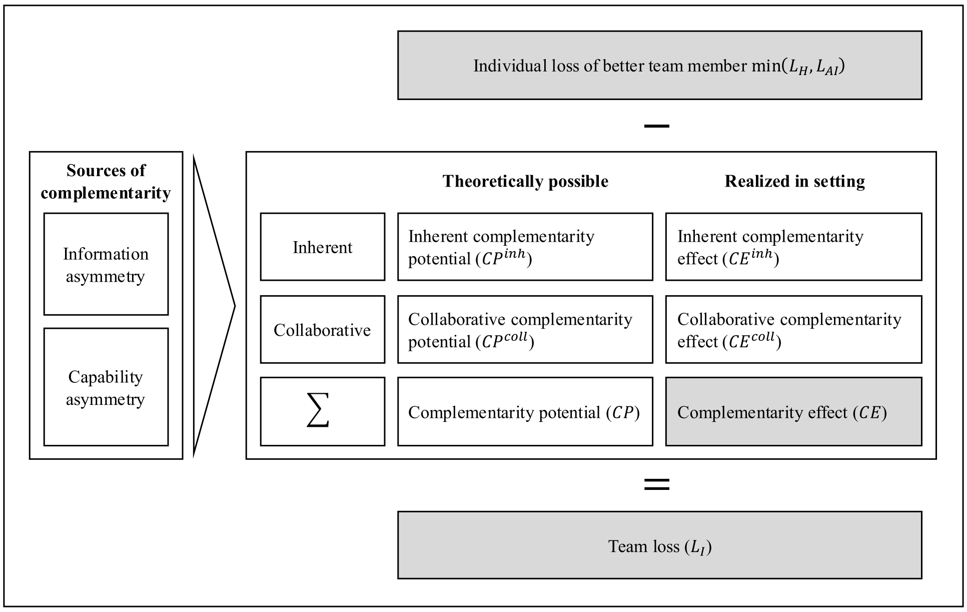

Individual loss of better team member min(L_H, L_{AI})
−
Sources of complementarity
Information asymmetry
Capability asymmetry
Theoretically possible
Realized in setting
Inherent
Inherent complementarity potential (CP^{inh})
Inherent complementarity effect (CE^{inh})
Collaborative
Collaborative complementarity potential (CP^{coll})
Collaborative complementarity effect (CE^{coll})
∑
Complementarity potential (CP)
Complementarity effect (CE)
=
Team loss (L_T)

Figure 6. Conceptual framework for human-AI complementarity. It summarizes the complementarity potential and effect, including the inherent and collaborative components. Information and capability asymmetry can affect both the theoretically existing potential and the realized effect.

individual human or AI decision—i.e., the human relies fully on his/her judgment or that of the AI—or it can be a function of the individual human and AI decision—i.e., an "integrated" decision that potentially differs from both individual ones. In the first experiment, we investigated the effect of *information asymmetry* on decision-making in the human-AI team in the form of additional contextual information only available to the human. In the second experiment, we investigated the effect of *capability asymmetry* on joint decision-making with different levels of diversity between the capabilities of humans and AI.

4.1. Experiment 1: the effect of information asymmetry

In the first experiment, we applied the conceptualization developed in Section 3 to study the effect of information asymmetry between humans and AI as a relevant source of complementarity. More precisely, we created an intervention in which humans are given contextual information withheld from the AI to investigate whether and how this affects the final team decision and the realized complementarity effect.

4.1.1. Task and AI model

We drew on a real estate appraisal task provided on the data science website kaggle.com (Kaggle, 2019). Since housing is a basic need, and because it is ubiquitous in everyone's life, all people to some degree have the ability to assess a house's value on the basis of relevant factors such as size or appearance. The data set encompasses 15,474 houses and contains information about the street, city, number of bedrooms, number of bathrooms, and size (in square feet). In the data set, the house prices denote their listing price. The average house price is $703,120, ranging between a minimum of $195,000 and a maximum of $2,000,000. An image of each house is also provided.

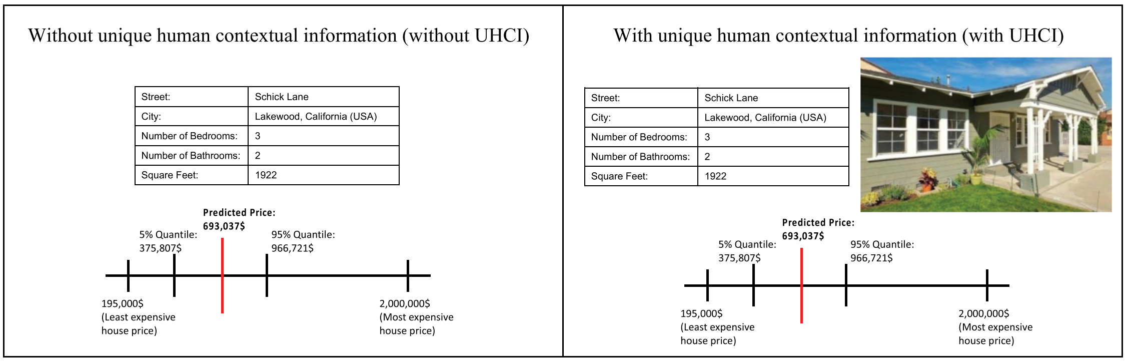
In respect of the house price prediction task, we implemented a random forest regression model as the AI model (Breiman, 2001). We drew on the individual trees in the random forest to generate a predictive distribution for each instance and provided the 5% and 95% quantiles as indicators of the AI model's prediction uncertainty. We used 80% of the data as the training set and 20% as the test set. We trained the random forest on the following features: the street, city, number of bedrooms, number of bathrooms, and square feet of the house. The house's image was withheld from the AI model.

In respect of the behavioral experiment, we focused on detached family houses in the test set, with the existing image providing a view of its exterior. From these, we randomly drew a hold-out set of 15 houses to serve as samples for our behavioral experiment. The AI model achieves a performance measured in terms of the mean absolute error (MAE) of $163,080 regarding the hold-out set, which is representative of its performance on the entire test set. In respect of the condition with unique human contextual information (UHCI), we gave humans an additional image of the house, which is likely to constitute valuable information. Humans are able to leverage their general understanding to form an overall assessment based on the house's features, the visible surroundings, and its appearance. We conducted an initial pilot study to verify this assumption (Appendix B.1 contains additional details).

4.1.2. Study design

We conducted an online experiment with a between-subject design. We recruited participants from prolific.com. The study included two conditions (one with UHCI and one without UHCI) and randomly assigned each participant to one of them. In the condition without UHCI, participants were only given the houses' tabular data, while participants in the condition with UHCI were also given images of the house. We did not allow any repeated participation. Each participant passed the following steps: First, they were asked for their consent and to answer a control question. After receiving instructions about the study, participants had to complete a tutorial to familiarize themselves with the task, the data and the AI. Subsequently, they conducted two training task instances before being transferred to the main task which consisted of 15 house price task instances presented in random order. For each instance, participants had to provide a prediction before receiving the AI's recommendation. Then, they were asked to adjust the AI's prediction in the best possible way, constituting the joint human-AI team prediction. Finally, participants had to complete a questionnaire regarding qualitative feedback and demographic information. Figure 7 depicts the main task's interface for both conditions. We refer to Appendix B.2 for a detailed description of the study design.

The overall task lasted approximately 30 minutes. Before recruiting participants, we computed the required sample size in a power analysis using G*Power (Faul et al., 2007). Based on the pilot data, we expected a large effect ($d = 0.8$). We referred to an alpha value of 0.05, while taking multiple testing into account in order to achieve a power of 0.8. This resulted in a total sample size of 86. Anticipating that some participants will fail the attention checks, we recruited a total of 120 participants (60 per condition). They received a base payment of £5 and are additionally incentivized following the approach of Kvaløy et al. (2015), who show the benefits of combining non-monetary motivators, such as recognition, attention, and verbal feedback, with performance-based pay. We



Without unique human contextual information (without UHCI)

Street:	Schick Lane
City:	Lakewood, California (USA)
Number of Bedrooms:	3
Number of Bathrooms:	2
Square Feet:	1922

With unique human contextual information (with UHCI)

Street:	Schick Lane
City:	Lakewood, California (USA)
Number of Bedrooms:	3
Number of Bathrooms:	2
Square Feet:	1922

Figure 7. An overview of the interfaces containing the information that the participants were given in the respective behavioral experiment's treatments.

achieved this by adding motivational statements and by giving the top 10% participants an additional pound. Note that the two training task instances are not included in the final evaluation. To ensure the quality of the collected data, we removed those participants whose entered prices exceed the communicated maximum house price of $2,000,000 in the data set. We also identified outliers for removal by using the median absolute deviation (Leys et al., 2013; Rousseeuw & Croux, 1993). After applying these criteria, we continued with the data of 101 participants across both conditions—53 in the treatment without UHCI and 48 in that with UHCI (see Appendix B.3 for additional details about the participants).

4.1.3. Evaluation measures

For each participant, we measured the loss of the human (l_H), the AI (l_{AI}), and the team decision (l_I) as the absolute error, and calculated the average over all task instances to receive the human (L_H), the AI (L_{AI}), and the team performance (L_I) corresponding to the mean absolute error (MAE). Furthermore, we calculated the complementarity potential's and complementarity effect's respective components as defined in Section 3. Finally, for each measure, we calculated the average over all the participants.

4.1.4. Results

In this section, we analyze the impact of unique human contextual information on the team performance, the complementarity potential, and the complementarity effect. We evaluate the results' significance by using the Student's T-test and the Mann-Whitney U-test, depending on whether the prerequisites have been fulfilled. We apply the Bonferroni correction and adjust the p-values accordingly. First, we focus on the impact of contextual information on performance, followed by an in-depth analysis of its impact on complementarity potential and effect.

Figure 8 displays the isolated human and joint human-AI performance for both conditions. It also includes the performance of the AI alone. We first evaluate the impact of unique human contextual information without any AI assistance. Participants in the treatment without UHCI achieve an MAE of \$251,282, while those in the treatment with UHCI yield an MAE of \$200,510—an improvement of \$50,772 (20.21%), which is significant ($d = 0.92$, $p < 0.001$, two-sample, two-tailed T-test). This result confirms the general usefulness of the provided house images from the human perspective.

Next, we evaluate the impact of unique human contextual information when the human is teamed with the AI. The team performance in the treatment without UHCI results in an MAE of \$160,095 versus an MAE of \$148,009 in the treatment with UHCI—an improvement of \$12,086 (7.55%), which is significant ($d = 0.59$, $p < 0.05$, two-sample, two-tailed T-test). In both treatments, the human-AI team outperforms the AI (MAE: \$163,080). Whereas the difference between the performance of the human-AI team and the performance of the AI alone is significant in the treatment with UHCI ($d = 0.68$, $p < 0.001$, one-sample, two-tailed T-test), the difference in the treatment without UHCI does not constitute a significant improvement ($d = 0.16$, $p = 1.0$, one-sample, two-tailed T-test).

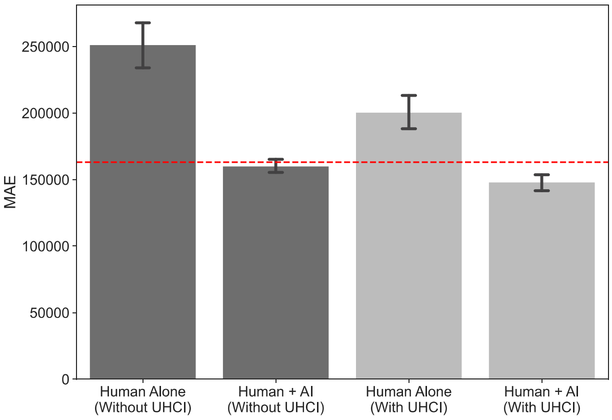


Figure 8. Performance results as the MAE across the conditions (UHCI = unique human contextual information), including 95% confidence intervals. The red horizontal line denotes the AI performance.

4.1.4.1. Complementarity potential. First, we analyze the inherent complementarity potential (CP^{inh}). We observe a significant increase due to the unique human contextual information. In the condition without UHCI, the CP^{inh} is \$42,995, and increases to \$61,970 in the condition with UHCI ($d = 1.05$, $p < 0.001$, two-tailed Mann-Whitney U test). This shows that the images contain useful contextual information for humans, which the AI cannot access, resulting in fewer shared errors that humans and AI make individually.

Next, we calculate the collaborative complementarity potential (CP^{coll}). Whereas in the condition without UHCI, the CP^{coll} results in \$120,085, in the condition with UHCI it amounts to \$101,110. The difference is statistically significant ($d = 1.05$, $p < 0.001$, two-tailed Mann-Whitney U test). Since the participants in both conditions work with the same AI model, which has an overall better individual performance, the CP (i.e., the sum of the inherent and collaborative component) is \$163,080 in both conditions. As CP^{inh} increases due to UHCI, CP^{coll} decreases because the CP remains constant.

4.1.4.2. Complementarity effect. Then, we focus on the realized complementarity potential, i.e., the complementarity effect (CE). We find a significant difference between the inherent complementarity effect (CE^{inh}) in both conditions (without UHCI: \$14,468; with UHCI: \$27,860; $d = 0.87$, $p < 0.001$, two-tailed Mann-Whitney U test), which highlights contextual information's potential. This absolute increase might be due to an increase in inherent complementarity potential and/or an improvement in the integration of both team members' predictions through the human. In order to investigate this further, we also calculate the inherent complementarity effect's ($\frac{CE^{inh}}{CP^{inh}}$) relative amount. This analysis reveals that unique human contextual information not only enhances the theoretically available inherent complementarity potential, but that the participants could also use significantly more of it (without UHCI: 34%; with UHCI: 45%; $d = 0.82$, $p < 0.001$, two-tailed Mann-Whitney U test).

Next, we analyze unique human contextual information's impact on the collaborative complementarity effect (CE^{coll}). We do not find a significant difference between the two treatments (without UHCI: \$–11,483; with UHCI: \$–12,789; $d = 0.08$, $p = 1.0$, two-tailed Mann-Whitney U test). The negative collaborative complementarity effect results from the AI outperforming its human team member individually in the experimental setup. A positive CE^{coll} can only occur on individual task instances where the team loss is even lower than that of the AI and the human alone (see Figure 5). Nevertheless, investigating the individual performance of humans and AI as well as the team performance for each house separately reveals that the human-AI team can derive team decisions for task instances 1, 4, and 13 that are, on average over all participants, more accurate compared to the respective individual decisions of human and AI. This demonstrates the occurrence of a positive collaborative complementarity effect for these three task instances. See Appendix B.4 for detailed results regarding a performance analysis of each house.

Finally, CE^{inh} and CE^{coll} can be summed to obtain the total complementarity effect (CE), which equals the performance difference between the best individual team member and the joint human-AI team performance (without UHCI: \$2,985; with UHCI: \$15,071). Figure 9 summarizes the results of our experiment.

4.2. Experiment 2: the effect of capability asymmetry

In the second behavioral experiment, we applied the conceptualization to study the effect of capability asymmetry between humans and AI as another relevant source of complementarity. In detail, starting with a "baseline" AI, we created an intervention in which we increase the asymmetry between the AI's and the humans' capabilities while keeping its overall performance constant. This means the second AI tends to make correct decision suggestions for task instances that tend to be more difficult for humans ("complementary" AI).

4.2.1. Task and AI model

To investigate the effect of capability asymmetry on the human-AI team performance in decision-making, we chose the image recognition context. Research has demonstrated that humans and AI tend to make different errors on image classification tasks (Fügener et al., 2021; Steyvers et al., 2022). Specifically, an AI model based on deep convolutional neural networks tends to infer classification decisions differently than humans do (Geirhos et al., 2020, 2021). We could therefore expect the AI model to classify certain images more accurately than humans and vice versa, thereby creating inherent complementarity potential. However, it remains unclear whether this naturally existing potential could be sufficiently realized when humans incorporate the AI decision into a final team decision, and whether its increase in the intervention affects the realization.

In order to undertake the experiment, we drew on the image data set that Steyvers et al. (2022) provided. The data set comprises 1,200 images distributed evenly across 16 classes (e.g., airplane, dog, or car). It was curated on the basis of the ImageNet Large Scale

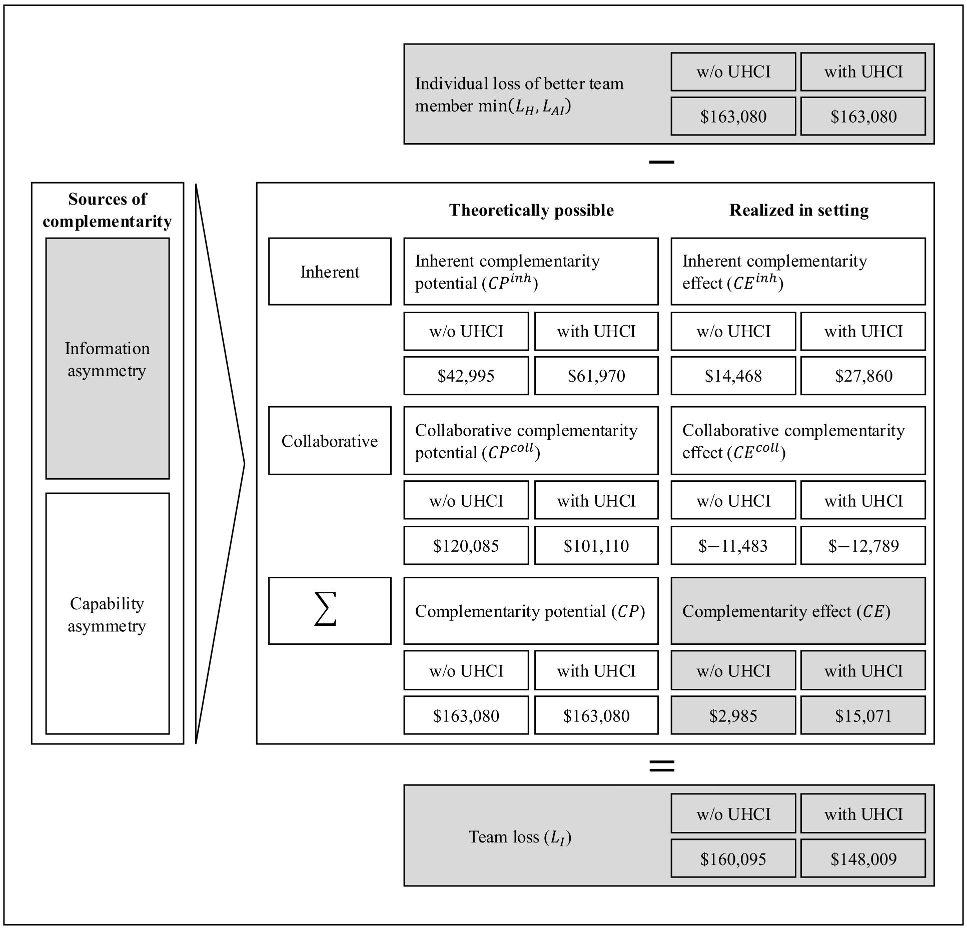


Figure 9. Result summary of the real estate appraisal experiment.

Visual Recognition Challenge (ILSVRC) 2012 database (Russakovsky et al., 2015). To increase the task difficulty for humans and the AI, the authors applied phase noise distortion at each spatial frequency, which was uniformly distributed in the interval $[-\omega, \omega]$ with $\omega = 110$. Despite the heightened difficulty level, both humans and AI can attain comparable performance on the task. In addition to ground truth labels, the data set also contains multiple human predictions for each image provided by crowd workers, allowing us to infer a proxy for human classification difficulty. Images with a high disagreement in respect of multiple human predictions indicate a higher level of difficulty for humans.

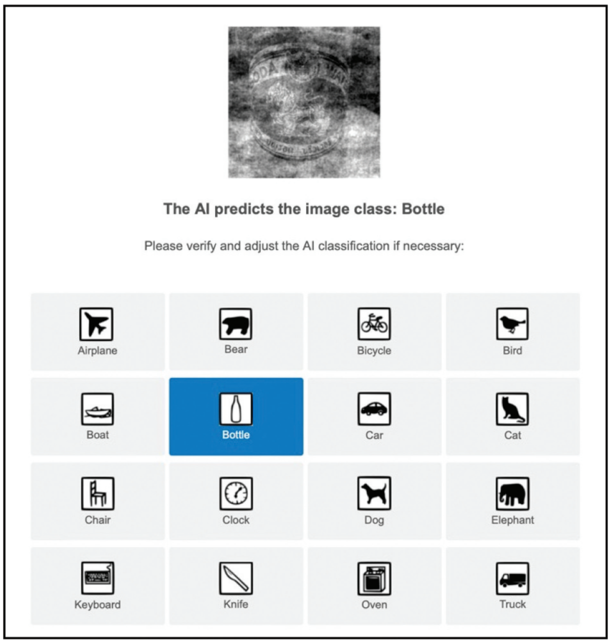
We implemented the AI model as a convolutional neural network, more precisely, as a DenseNet161 (Huang et al., 2017), pre-trained on ImageNet. We partitioned the data set into a training (60%), validation (20%), and test set (20%). In the baseline condition, we fine-tuned the AI model on the distorted images over 100 epochs, applying early stopping on the validation loss. We used SGD as an optimizer with a learning rate of $1{\cdot}10^{-4}$, a weight decay of $5{\cdot}10^{-4}$, a cosine annealing learning rate scheduler, and a batch size of 32. The AI model achieves a classification error of 26.66% on the test set.

In the intervention, we created an alternative AI model that makes erroneous decisions for different instances. We fine-tuned the DenseNet161 model exactly as in the baseline condition, but, for each image in the training set, also incorporated a human prediction as an additional label in the training process in order to incentivize the AI model to learn to correctly classify the images that tend to be more difficult for humans (Hemmer et al., 2022; Madras et al., 2018; Wilder et al., 2020). See Appendix C.1 for additional implementation details of this approach. Even though the AI model has a slightly higher classification error of 33.75%, this approach results in higher capability asymmetry, i.e., non-overlapping capabilities between the humans and the AI model. We selected 15 images from the test set for the experiment, such that both AI models exhibit the same performance on the test set (26.66%), while considering non-overlapping errors between both AI models.

4.2.2. Study design

We conducted an online experiment, employing a between-subject design with two conditions. Participants were recruited from prolific.com and randomly assigned to one of the conditions; repeated participation was not allowed. We employed a similar experimental set up as in the first experiment: Participants were transferred to the experimental website where they had to submit a consent form and answer a control question and an attention check. Subsequently, they received a tutorial about the task, the data, and the AI followed by a practice round comprising three images which had to be classified without AI assistance. Next, in the main task, participants had to classify 15 images in randomized order. First, they had to provide their own classification before receiving AI advice and were subsequently asked to verify and adjust the AI classification if necessary. Finally, participants had to answer a questionnaire regarding qualitative feedback and demographic information. Figure 10 displays the interface of the main task. We refer to Appendix C.2 for further information on the study design.

The overall task lasted approximately 20 minutes. Before recruiting participants, we computed the required sample size in a power analysis using G*Power (Faul et al., 2007). We tested for a medium to large effect ($d = 0.65$) and considered an alpha value of 0.05, while taking multiple testing into account in order to achieve a power of 0.8. This resulted in a total sample size of 128. In order to buffer for participants potentially failing attention checks, we recruited a total of 170 participants. They received a base payment of £8 and were additionally incentivized following the approach pursued in the first behavioral experiment (Kvaløy et al., 2015). We excluded participants who did not pass the integrated attention checks, resulting in 144 participants—76 in the base condition (baseline AI) and 68 in the intervention (complementary AI). We provide further details in Appendix C.3.



The AI predicts the image class: Bottle

Please verify and adjust the AI classification if necessary:

Figure 10. An overview of the interfaces that the participants are shown in both treatments of the behavioral experiment.

4.2.3. Evaluation measures

For each participant, the loss of the human (l_H), the AI (l_{AI}), and the team decision (l_I) was measured as the classification error and averaged over all the task instances, providing the human (L_H), the AI (L_{AI}), and the team performance (L_I). We also calculated the complementarity potential and effect, including their components (see Section 3). Finally, for each measure, we calculated the average over all the participants.

4.2.4. Results

We analyze the effect of capability asymmetry on team performance, complementarity potential, and complementarity effect while assessing the statistical significance by using the same procedure as in the first experiment.

Figure 11 shows the classification error for humans performing the task alone and together with the AI in both conditions. In addition, it also includes the classification error of both AI models, which are identical in this task. Humans conducting the task alone exhibit a classification error of approximately 0.30, which is nearly identical across the conditions (Baseline AI: 0.2999; Complementary AI: 0.2951; $d = 0.05$, $p = 1.0$, two-sample, two-tailed T-test). When humans are teamed with the AI, the joint performance increases in both conditions. Whereas the human-AI team yields a classification error of 0.2473 in the condition with the baseline AI, this error decreases even further to 0.1461 in the team with the complementary AI. This corresponds to an

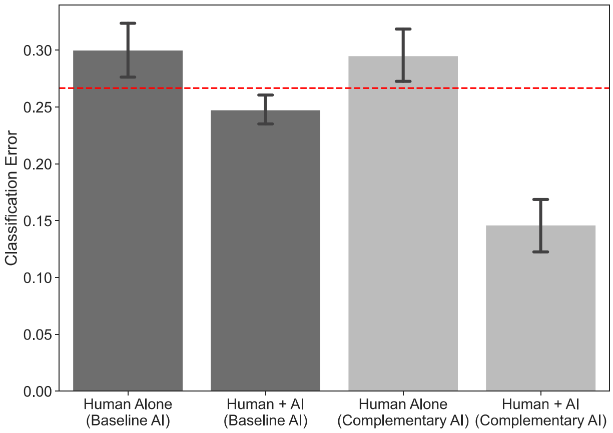


Figure 11. Performance results as classification error across conditions, including 95% confidence intervals. The red horizontal line denotes the AI performance.

improvement of 41%, which is significant ($d = 1.29$, $p < 0.001$, two-sample, two-tailed T-test). Both classification errors are significantly lower than that of the AI conducting the task alone in both conditions (Baseline AI: 0.2666, $d = 0.33$, $p < 0.05$, one-sample, two-tailed T-test; Complementary AI: 0.2666, $d = 1.25$, $p < 0.001$, one-sample, two-tailed T-test).

4.2.4.1. Complementarity potential. We observe a significant increase in the inherent complementarity potential (CP^{inh}) in the condition with the complementary AI (Baseline AI: 0.0640, Complementary AI: 0.2480; $d = 3.81$, $p < 0.001$, two-tailed Mann-Whitney U test). This reflects a higher level of capability asymmetry between humans and the AI. Compared to the baseline AI condition, the AI makes more erroneous decisions for instances that humans can process correctly, whereas humans err for instances that AI can classify correctly. Conversely, we observe a significant decrease in the collaborative complementarity potential (CP^{coll}) in the condition with the complementary AI (Baseline AI: 0.2026, Complementary AI: 0.0186, $d = 3.81$, $p < 0.001$, two-tailed Mann-Whitney U test) as the CP remains constant due to the AI being individually more accurate than the humans. Whereas the inherent complementarity potential constitutes 24% of the overall complementarity potential in the baseline condition, this share rises to 93% in the complementary AI condition. This means that, for the majority of the task instances, one team member is theoretically capable of making a correct decision.

4.2.4.2. Complementarity effect. In the baseline condition, 58% of the inherent complementarity potential ($\frac{CE^{inh}}{CP^{inh}}$) could be realized by the humans integrating their own decision and that of the AI into a final team decision, resulting in a CE^{inh} of 0.0368. In the condition with the complementary AI, it is possible to realize 89% of the inherent complementarity potential, resulting in a CE^{inh} of 0.2196. This shows a significant performance improvement ($d = 3.43$, $p < 0.001$, two-tailed Mann-Whitney U test), which is attributable to a significantly larger fraction of $\frac{CE^{inh}}{CP^{inh}}$ that could be realized ($d = 1.02$, $p < 0.001$, two-tailed Mann-Whitney U test). It indicates that humans tended to rely on the AI decisions when they were correct, but on their decision when it was incorrect. Moreover, in both conditions the collaborative complementarity effect (CE^{coll}) is negative. Whereas the value is only slightly negative in the baseline condition (Baseline AI: −0.0175), it decreases to −0.0990 in the condition with the complementary AI. The difference between the two conditions is significant ($d = 1.32$, $p < 0.001$, two-tailed Mann-Whitney U test). In addition, we refer to Appendix C.4 for additional analyses, including an analysis of each image. In this context, similar to the first behavioral experiment, we find a positive collaborative complementarity effect which occurs for individual participants for three task instances.

In summary, CE^{inh} and CE^{coll} result in the total complementarity effect (CE)—equivalent to the performance difference between the best individual team member and the team performance (Baseline AI: 0.0193, Complementary AI: 0.1206). Figure 12 summarizes the analysis.

5. Discussion

We begin this section with a discussion of the key findings of this work. We then highlight how our research advances the IS community’s theoretical understanding of human-AI collaboration in decision-making before discussing its managerial implications. Finally, we outline limitations and consider potential areas for future research before concluding this work.

5.1. Key findings

In this study, we contribute to a comprehensive understanding of the inner workings of human-AI teams in the context of decision-making, and provide support on how to achieve CTP more consistently. In particular, we address our research questions to make the following contributions: First, we develop a conceptualization of human-AI complementarity that introduces and formalizes the notions of complementarity potential and complementarity effect (RQ1). Second, we identify and outline information and capability asymmetries as sources of complementarity (RQ2). Third, we provide empirical evidence for the utility of our conceptualization in two behavioral experiments that individually demonstrate the relevance of each of the two sources (RQ1 and RQ2).

The first experiment highlights that equipping humans with unique contextual *information* can not only enlarge the inherent complementarity potential, but may also disproportionally increase the realized gain, i.e., the inherent complementarity effect, and, thus, materialize in CTP. This constitutes an interesting finding as, intuitively, the human perception of having more information than the AI could have led to a decreased reliance on AI suggestions in team predictions. As a consequence, this algorithm aversion could have left existing potential for performance improvement untapped (Jussupow et al., 2020; Longoni et al., 2019; Mahmud et al., 2022; Sieck & Arkes, 2005).

The second experiment reveals that a larger *capability* asymmetry can help capture a larger share of inherent complementarity potential, resulting in CTP. This is also an insightful observation, as we could have expected that humans may reject even correct AI

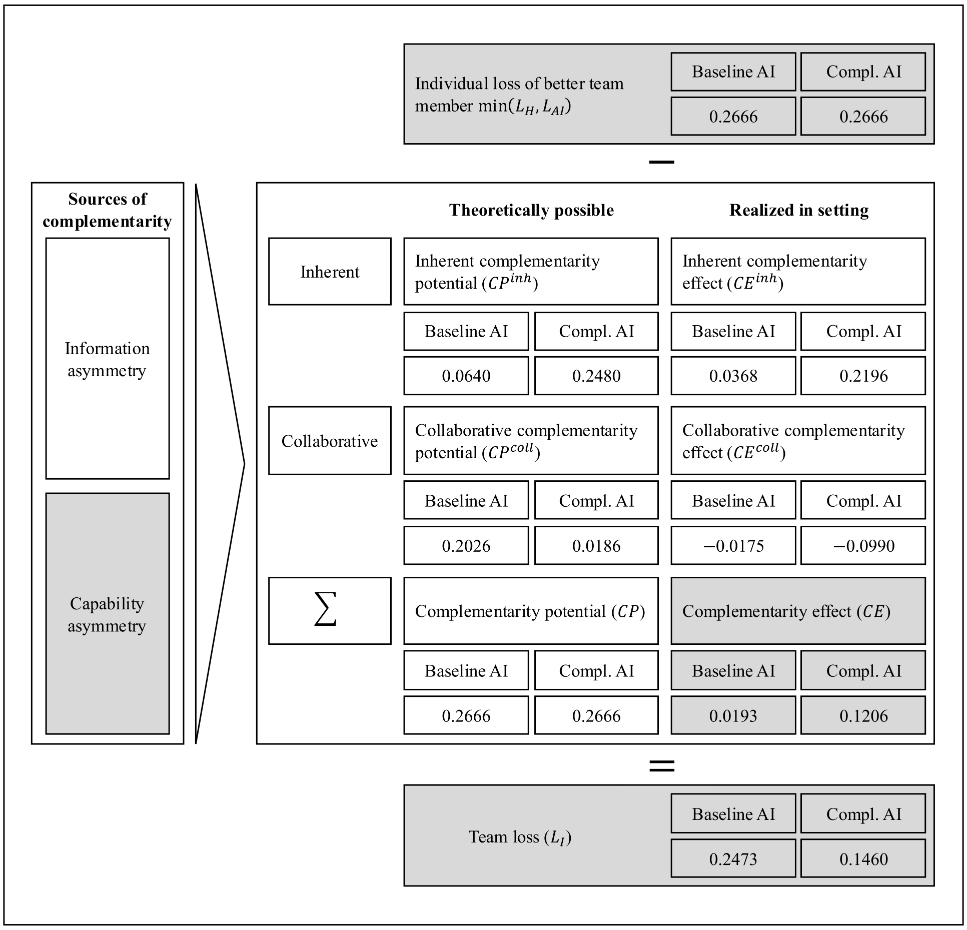


Figure 12. Result summary of the image classification experiment.

assistance once they witness the AI committing errors on task instances they can easily solve themselves (Dietvorst et al., 2015). The decrease in the collaborative complementarity effect reveals that for some instances humans either rely on incorrect AI decisions or do not rely on correct ones. However, this is overcompensated by the increase in the inherent complementarity effect. Overall, this finding contributes to our understanding of the impact of diversity in team composition on performance (Horwitz, 2005). We know from human teams that expertise diversity can be both a performance driver—e.g., as it provides a broader range of cognitive skills (Cohen & Bailey, 1997)—as well as a performance inhibitor due to the potential difficulties of achieving a mutual understanding (Dougherty, 1992). Similarly, there may also be a trade-off in human-AI teams, as humans could develop algorithm aversion and avoid AI suggestions if they observe them to be incorrect too frequently (Dietvorst et al., 2015). Our conceptualization enables a detailed analysis of this trade-off—finding that, in the specific experiment, the capability asymmetry between humans and AI affects team performance positively.

Finally, the results provide empirical evidence for the validity of our conceptualization. In particular, we find that not only are diverse human-AI teams able to realize the inherent complementarity potential, but also that the interaction between humans and AI can lead to different and more accurate team decisions for certain task instances compared to their individual decisions. This proves the existence of the collaborative component of the complementarity potential over and above the inherent one.

5.2. Theoretical contributions and implications

Our research contributes to the IS literature by advancing the theoretical understanding of human-AI complementarity in decision-making. With our work, we contribute to the ongoing academic discourse on augmenting human capabilities with AI. In this context, our contributions underline the "AI *with* human" perspective (Huysman, 2020), thereby not only advancing the research community's theoretical understanding of human-AI teams, but also providing empirical evidence that the combination of human and AI capabilities can actually result in superior decision-making outcomes.

Moreover, we extend the body of knowledge by proposing a conceptualization together with a novel measurement concept for complementarity potential and effect. In particular, our conceptualization allows researchers to systematically assess *when* and, more importantly, *how* human-AI teams can achieve CTP as it enables a more nuanced and measurable analysis of their synergetic potential. This can support researchers in formulating and testing hypotheses for design theories for human-AI teams in the future (Jain et al., 2021). While previous studies have confined themselves to measure aggregated performance metrics that do not provide insights into the inner workings of the team (Hemmer et al., 2021), our conceptualization fosters a profound understanding of human-AI decision-making by differentiating between the inherent and collaborative components.

In addition, the identified sources of complementarity constitute concrete factors that researchers should take into consideration when developing design principles for human-AI collaboration—underscored by the empirical evidence of our behavioral experiments. Current research efforts often focus on developing design features of the AI, e.g., explanations (Liu et al., 2021; van der Waa et al., 2021), to improve the decision quality of the human. This work promotes another avenue for improving the outcomes of human-AI collaboration by focusing on the complementary composition of human-AI teams.

Finally, our empirical results also inform the discourse on algorithmic aversion. Previous research has investigated factors and conditions that foster humans' mistrust in algorithmic decisions and can lead to inferior collaborative outcomes (Dietvorst et al., 2015; Jussupow et al., 2020; Mahmud et al., 2022). Our empirical results demonstrate that, in our specific experiments, leveraging information and capability asymmetries does not lead humans to avoid AI advice despite having more information and witnessing AI errors for task instances that are relatively easy for them. These insights emphasize the importance of utilizing these sources of complementarity to improve human-AI collaboration in decision-making.

5.3. Managerial implications

Our work has also important implications for managerial decision-makers. First and foremost, it makes a compelling case for the purposeful integration of humans and AI. In particular, it should redirect mere automation discussions and rather guide practitioners towards designing human-AI collaboration in ways that reap synergistic benefits. Second, the work identifies information and capability asymmetries as the key sources of complementarity: Thoughtful analyses of the capabilities of humans and AI with respect to these sources should point practitioners to promising use cases for human-AI collaboration across various application domains. In particular, this implies to not build AI models that mimic human experts but rather to target complementary capabilities, i.e., by training AI models excelling for task instances where human decision-making fails (Hemmer et al., 2022; Mozannar & Sontag, 2020; Wilder et al., 2020). From a human perspective, this also means investing in the development and preservation of human knowledge that is not covered by AI capabilities (Spitzer et al., 2023). Third, as demonstrated in both experiments, the conceptual framework (Figure 6) can be applied to quantify complementarity effects and to plan and monitor the results of human-AI teams. With these foundations, practitioners can explore various design options to build and engage human-AI teams for complementarity, including the composition of suitably diverse teams and the design of appropriate collaboration mechanisms.

A simple example may demonstrate these implications: Consider the X-ray analysis performed by radiologists in a hospital. AI-based interpretation of X-rays should not be applied to render human experts obsolete but rather to add AI support in cases where the AI can draw on additional information (e.g., access to a broader number of cases) or enjoys capability advantages (e.g., higher analysis speed). Practitioners should ensure that the respective strengths of the team members are aligned when assembling human-AI teams and that the collaboration mechanisms are conducive to realizing the complementarity potential. The effectiveness of the devised setup can be measured (and continuously monitored) using the developed framework, as shown in Figures 9 and 12.

We anticipate that both the notion of complementarity as well as the operational support presented in this work will massively influence the way human-AI teams are built and coordinated.

5.4. Limitations and directions for future research

Our current research has several limitations that at the same time open avenues for further research: First, the controlled settings of our laboratory-based behavioral experiments allow us to derive the insights presented in this work. However, they do not yet consider the *wider range of complex human factors* that shape the effectiveness of human-AI teams, such as motivation (Schunk, 1995), engagement (Benz et al., 2024; Chandra et al., 2022), self-efficacy (Westphal et al., 2024), behavioral patterns within teams (Schecter et al., 2022), situational factors like time pressure (Cao et al., 2023; Swaroop et al., 2024), or acceptance barriers of AI systems (Jain et al., 2021).

Second, we use a sequential decision-making setup to first measure the human decision and then the team decision. *Different collaboration forms and mechanisms* that do not ex ante reveal the AI decision (Green & Chen, 2019) or do not integrate a team decision, but simply delegate task instances within the human-AI team (Fügener et al., 2022) may produce different results. Future work should investigate these collaboration forms as to their suitability for specific use cases and application domains.

Third, we have described information and capability asymmetry as the key sources of complementarity, and investigated them in behavioral experiments. A deeper *exploration of complementarity sources* could reveal and categorize different forms of asymmetry and develop metrics to measure them quantitatively.

Fourth, we have considered a situation with one AI and one human. However, team settings could comprise more than two team members. *Team design principles* could be derived to ensure a sufficient degree of complementarity and to appropriately select and combine multiple human and artificial team members (Hemmer et al., 2022).

Finally, although human-AI decision-making is an important application of human-AI collaboration, *other types of problems*, e.g., creative or generative tasks (Schmidt et al., 2023) may also receive scrutiny with regard to complementarity (Vaccaro et al., 2024).

5.5. Conclusion

So far, human-AI collaboration in decision-making has been primarily concerned with AI systems helping human users. However, since the number of decision tasks that can be automated (i.e., can be solved by the AI alone) is increasing steadily, the focus has shifted to the purposeful design of the collaboration between humans and AI as team members—thereby shaping the future of work with AI. The ultimate objective of these teams must be to achieve complementary team performance (CTP), with the team outperforming each individual team member. The IS community is predestined to drive the development of appropriate theories and to lay the foundation for practical applications. We hope that the conceptual foundation developed in this paper will provide fruitful ground for future research, and that the empirical studies illustrate the validity and potential of the human-AI complementarity paradigm.

Notes

1. Throughout this paper, we refer to models or systems using Machine Learning (ML) as “Artificial Intelligence” (AI) (Berente et al., 2021; Collins et al., 2021; Rai et al., 2019). While we acknowledge the technical distinction between AI and ML as discussed, e.g., by Kühl, Schemmer et al. (2022) and although we have considered the more precise reference to an “ML” model, we adopt the use of the broader “AI” term as the prevalent terminology established in the Computer Science and Human-Computer Interaction communities. This choice reflects the contemporary linguistic trend rather than a lack of distinction between the two fields.
2. We note that this is not possible if team members rely on one of their individual decisions as team decision —as task solutions are confined to the solutions provided by each team member alone.
3. For simplification, we report L_D in the introductory example as the sum instead of the average of all the instances.

Acknowledgements

We would like to express our sincere gratitude to the Associate Editor and the anonymous reviewers for their invaluable feedback and constructive suggestions throughout the review process. Their deep engagement with our work and thoughtful comments have significantly enhanced the quality and clarity of this paper. We are confident that the final manuscript is fundamentally better because of their contributions, which we highly appreciate.

Disclosure statement

No potential conflict of interest was reported by the authors. Please note this article does not represent an official IBM statement.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *Institute of Electrical and Electronics Engineers Access*, 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Afshar, P., Rafiee, M. J., Naderkhani, F., Heidarian, S., Enshaei, N., Oikonomou, A., Babaki Fard, F., Anconina, R., Farahani, K., Plataniotis, K. N., & Mohammadi, A. (2022). Human-level COVID-19 diagnosis from low-dose CT scans using a two-stage time-distributed capsule network. *Scientific Reports*, *12*(1), 4827–4838. https://doi.org/10.1038/s41598-022-08796-8

Alufaisan, Y., Marusich, L. R., Bakdash, J. Z., Zhou, Y., & Kantarcioglu, M. (2021). Does explainable artificial intelligence improve human decision-making? *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(8), 6618–6626. https://doi.org/10.1609/aaai.v35i8.16819

Ancona, D. G., & Caldwell, D. F. (1992). Demography and design: Predictors of new product team performance. *Organization Science*, *3*(3), 321–341. https://doi.org/10.1287/orsc.3.3.321

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). https://doi.org/10.1145/3411764.3445717

Bauer, K., von Zahn, M., & Hinz, O. (2023). Expl(ai)ned: The impact of explainable artificial intelligence on users' information processing. *Information Systems Research*, *34*(4), 1582–1602. https://doi.org/10.1287/isre.2023.1199

Benz, C., Riefle, L., & Satzger, G. (2024). User engagement and beyond: A conceptual framework for engagement in information systems research. *Communications of the Association for Information Systems*, *54*(1), 331–359. https://doi.org/10.17705/1CAIS.05412

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Special issue editor's comments: Managing artificial intelligence. *Management Information Systems Quarterly*, *45* (3), 1433–1450.

Breiman, L. (2001). Random forests. *Machine Learning*, *45* (1), 5–32. https://doi.org/10.1023/A:1010933404324

Brynjolfsson, E., Mitchell, T., & Rock, D. (2018). What can machines learn and what does it mean for occupations and the economy? *AEA Papers and Proceedings*, Philadelphia, USA (Vol. 108, pp. 43–47). https://doi.org/10.1257/pandp.20181019

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., & Lundberg, S. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv Preprint arXiv*, 1–155. https://doi.org/10.48550/arXiv.2303.12712

Buçinca, Z., Lin, P., Gajos, K. Z., & Glassman, E. L. (2020). Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. *Proceedings of the 25th International Conference on Intelligent User Interfaces*, Cagliari, Italy (pp. 454–464). https://doi.org/10.1145/3377325.3377498

Cambridge Dictionary. (2024). *English dictionary*. Retrieved August 2, 2024. https://dictionary.cambridge.org/us/dictionary/english/complementarity

Cao, S., Gomez, C., & Huang, C.-M. (2023). How time pressure in different phases of decision-making influences human-AI collaboration. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW2), 1–26. https://doi.org/10.1145/3610068

Carton, S., Mei, Q., & Resnick, P. (2020). Feature-based explanations don't help people detect misclassifications of online toxicity. *Proceedings of the International AAAI Conference on Web & Social Media*, *14*(1), 95–106. https://doi.org/10.1609/icwsm.v14i1.7282

Chandra, S., Shirish, A., & Srivastava, S. C. (2022). To be or not to be . . .human? Theorizing the role of human-like competencies in conversational artificial intelligence agents. *Journal of Management Information Systems*, *39* (4), 969–1005. https://doi.org/10.1080/07421222.2022.2127441

Cohen, S. G., & Bailey, D. E. (1997). What makes teams work: Group effectiveness research from the shop floor to the executive suite. *Journal of Management*, *23*(3), 239–290. https://doi.org/10.1177/014920639702300303

Collins, C., Dennehy, D., Conboy, K., & Mikalef, P. (2021). Artificial intelligence in information systems research: A systematic literature review and research agenda. *International Journal of Information Management*, *60*, 1–17. https://doi.org/10.1016/j.ijinfomgt.2021.102383

D'Arcy, J., Gupta, A., Tarafdar, M., & Turel, O. (2014). Reflecting on the "dark side" of information technology use. *Communications of the Association for Information Systems*, *35*(1), 5. https://doi.org/10.17705/1CAIS.03505

Day, M.-Y., Cheng, T.-K., & Li, J.-G. (2018). AI robo-advisor with big data analytics for financial services. *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, Barcelona, Spain (pp. 1027–1031). https://doi.org/10.1109/ASONAM.2018.8508854

Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid intelligence. *Business & Information Systems Engineering*, *61*(5), 637–643. https://doi.org/10.1007/s12599-019-00595-2

de Véricourt, F., & Gurkan, H. (2023). Is your machine better than you? You may never know. *Management Science*, 1–17. https://doi.org/10.1287/mnsc.2023.4791

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Donahue, K., Chouldechova, A., & Kenthapadi, K. (2022). Human-algorithm collaboration: Achieving complementarity and avoiding unfairness. *Proceedings of the 2022 Conference on Fairness, Accountability, and Transparency*, Seoul, South Korea (pp. 1639–1656). https://doi.org/10.1145/3531146.3533221

Dougherty, D. (1992). Interpretive barriers to successful product innovation in large firms. *Organization Science*, *3*(2), 179–202. https://doi.org/10.1287/orsc.3.2.179

Dvijotham, K., Winkens, J., Barsbey, M., Ghaisas, S., Stanforth, R., Pawlowski, N., Strachan, P., Ahmed, Z., Azizi, S., Bachrach, Y., Culp, L., Daswani, M., Freyberg, J., Kelly, C., Kiraly, A., Kohlberger, T., McKinney, S., Mustafa, B. , and Cemgil, T. (2023). Enhancing the reliability and accuracy of AI-enabled diagnosis via complementarity-driven deferral to clinicians. *Nature Medicine*, *29*(7), 1814–1820. https://doi.org/10.1038/s41591-023-02437-x

Endsley, M. R. (2023). Supporting human-AI teams: Transparency, explainability, and situation awareness. *Computers in Human Behavior*, *140*, 1–16. https://doi.org/10.1016/j.chb.2022.107574

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Findling, C., & Wyart, V. (2021). Computation noise in human learning and decision-making: Origin, impact, function. *Current Opinion in Behavioral Sciences*, *38*, 124–132. https://doi.org/10.1016/j.cobeha.2021.02.018

Förster, M., Broder, H. R., Fahr, M. C., Klier, M., & Fink, L. (2024). Tell me more, tell me more: The impact of explanations on learning from feedback provided by artificial intelligence. *European Journal of Information Systems*, 1–23. https://doi.org/10.1080/0960085X.2024.2404028

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2021). Will humans-in-the-loop become borgs? Merits and pitfalls of working with AI. *Management Information Systems Quarterly*, *45*(3), 1527–1556. https://doi.org/10.25300/MISQ/2021/16553

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2022). Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation. *Information Systems Research*, *33*(2), 678–696. https://doi.org/10.1287/isre.2021.1079

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, *2*(11), 665–673. https://doi.org/10.1038/s42256-020-00257-z

Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A., & Brendel, W. (2021). Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*, *34*, 23885–23899.

Gerber, A., Derckx, P., Döppner, D. A., & Schoder, D. (2020). Conceptualization of the human-machine symbiosis - a literature review. *Proceedings of the 53rd Hawaii International Conference on System Sciences*, Grand Wailea, USA (Vol. 3. pp. 289–298).

Gladstein, D. L. (1984). Groups in context: A model of task group effectiveness. *Administrative Science Quarterly*, *29* (4), 499–517. https://doi.org/10.2307/2392936

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, bayesian learning mechanisms, and the theory theory. *Psychological Bulletin*, *138* (6), 1085–1108. https://doi.org/10.1037/a0028044

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). When will AI exceed human performance? Evidence from AI experts. *The Journal of Artificial Intelligence Research*, *62*, 729–754. https://doi.org/10.1613/jair.1.11222

Grace, K., Stewart, H., Sandkühler, J. F., Thomas, S., Weinstein-Raun, B., & Brauner, J. (2024). Thousands of AI authors on the future of AI. *arXiv Preprint arXiv*, 1–38. https://doi.org/10.48550/arXiv.2401.02843

Green, B., & Chen, Y. (2019). The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–24. https://doi.org/10.1145/3359152

Grønsund, T., & Aanestad, M. (2020). Augmenting the algorithm: Emerging human-in-the-loop work configurations. *Journal of Strategic Information Systems*, *29*(2), 1–16. https://doi.org/10.1016/j.jsis.2020.101614

Hackman, J. R., & Morris, C. G. (1975). Group tasks, group interaction process, and group performance effectiveness: A review and proposed integration. *Advances in Experimental Social Psychology*, *8*, 45–99. https://doi.org/10.1016/S0065-2601(08)60248-8

He, G., Kuiper, L., & Gadiraju, U. (2023). Knowing about knowing: An illusion of human competence can hinder appropriate reliance on AI systems. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany (pp. 1–18). https://doi.org/10.1145/3544548.3581025

Hemmer, P., Schellhammer, S., Vössing, M., Jakubik, J., & Satzger, G. (2022). Forming effective human-AI teams: Building machine learning models that complement the capabilities of multiple experts. *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, Vienna, Austria (pp. 2478–2484). https://doi.org/10.24963/ijcai.2022/344

Hemmer, P., Schemmer, M., Vössing, M., & Kühl, N. (2021). Human-AI complementarity in hybrid intelligence systems: A structured literature review. *Proceedings of the 25th Pacific Asia Conference on Information Systems*, Dubai, UAE (pp. 1–14).

Herath Pathirannehelage, S., Shrestha, Y. R., & von Krogh, G. (2024). Design principles for artificial intelligence-augmented decision making: An action design research study. *European Journal of Information Systems*, 1–23. https://doi.org/10.1080/0960085X.2024.2330402

Horwitz, S. K. (2005). The compositional impact of team diversity on performance: Theoretical considerations. *Human Resource Development Review*, *4*(2), 219–245. https://doi.org/10.1177/1534484305275847

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Honolulu, USA (pp. 4700–4708). https://doi.org/10.1109/CVPR.2017.243

Huysman, M. (2020). Information systems research on artificial intelligence and work: A commentary on “Robo-Apocalypse cancelled? Reframing the automation and future of work debate”. *Journal of Information Technology*, *35*(4), 307–309. https://doi.org/10.1177/0268396220926511

Ibrahim, R., Kim, S.-H., & Tong, J. (2021). Eliciting human judgment for prediction algorithms. *Management Science*, *67*(4), 2314–2325. https://doi.org/10.1287/mnsc.2020.3856

Inkpen, K., Chappidi, S., Mallari, K., Nushi, B., Ramesh, D., Michelucci, P., Mandava, V., Vepřek, L. H., & Quinn, G. (2023). Advancing human-AI complementarity: The impact of user expertise and algorithmic tuning on joint decision making. *ACM Transactions on Computer-Human Interaction*, *30*(5), 1–29. https://doi.org/10.1145/3534561

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., & Ng, A. Y. (2019). CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, Honolulu, USA (Vol. 33, pp. 590–597). https://doi.org/10.1609/aaai.v33i01.3301590

Jain, H., Padmanabhan, B., Pavlou, P. A., & Raghu, T. S. (2021). Editorial for the special section on humans, algorithms, and augmented intelligence: The future of work, organizations, and society. *Information Systems Research*, *32*(3), 675–687. https://doi.org/10.1287/isre.2021.1046

Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. *Proceedings of the 28th European Conference on Information Systems*, Marrakech, Morocco (pp. 1–16).

Jussupow, E., Spohrer, K., Heinzl, A., & Gawlitza, J. (2021). Augmenting medical diagnosis decisions? An investigation into physicians’ decision-making process with artificial intelligence. *Information Systems Research*, *32*(3), 713–735. https://doi.org/10.1287/isre.2020.0980

Kaggle. (2019). *House prices and images - SoCal*. Retrieved June 16, 2021. https://www.kaggle.com/ted8080/house-prices-and-images-socal

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *Quarterly Journal of Economics*, *133*(1), 237–293. https://doi.org/10.1093/qje/qjx032

Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: Review, synthesis, and future research directions. *European Journal of Information Systems*, *31*(3), 388–409. https://doi.org/10.1080/0960085X.2021.1927212

Kühl, N., Goutier, M., Baier, L., Wolff, C., & Martin, D. (2022). Human vs. supervised machine learning: Who learns patterns faster? *Cognitive Systems Research*, *76*, 78–92. https://doi.org/10.1016/j.cogsys.2022.09.002

Kühl, N., Schemmer, M., Goutier, M., & Satzger, G. (2022). Artificial intelligence and machine learning. *Electronic Markets*, *32*(4), 2235–2244. https://doi.org/10.1007/s12525-022-00598-0

Kunkel, J., Donkers, T., Michael, L., Barbu, C.-M., & Ziegler, J. (2019). Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow, Scotland (pp. 1–12). https://doi.org/10.1145/3290605.3300717

Kvaløy, O., Nieken, P., & Schöttner, A. (2015). Hidden benefits of reward: A field experiment on motivation and monetary incentives. *European Economic Review*, *76*, 188–199. https://doi.org/10.1016/j.euroecorev.2015.03.003

Lai, V., Chen, C., Smith-Renner, A., Liao, Q. V., & Tan, C. (2023). Towards a science of human-AI decision making: An overview of design space in empirical human-subject studies. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, Chicago, USA (pp. 1369–1385). https://doi.org/10.1145/3593013.3594087

Lai, V., Liu, H., & Tan, C. (2020). Why is "Chicago" deceptive? Towards building model-driven tutorials for humans. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu, USA (pp. 1–13). https://doi.org/10.1145/3313831.3376873

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, 1–72. https://doi.org/10.1017/S0140525X16001837

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*(4), 764–766. https://doi.org/10.1016/j.jesp.2013.03.013

Li, Y., & Ibanez-Guzman, J. (2020). Lidar for autonomous driving: The principles, challenges, and trends for automotive lidar and perception systems. *IEEE Signal Processing Magazine*, *37*(4), 50–61. https://doi.org/10.1109/MSP.2020.2973615

Licklider, J. C. R. (1960). Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics*, *HFE-1* (1), 4–11. https://doi.org/10.1109/THFE2.1960.4503259

Liu, H., Lai, V., & Tan, C. (2021). Understanding the effect of out-of-distribution examples and interactive explanations on human-AI decision making. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–45. https://doi.org/10.1145/3479552

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, *46*(4), 629–650. https://doi.org/10.1093/jcr/ucz013

Lu, T., & Zhang, Y. (2024). 1 + 1 > 2? Information, humans, and machines. *Information Systems Research*. https://doi.org/10.1287/isre.2023.0305

Ma, S., Lei, Y., Wang, X., Zheng, C., Shi, C., Yin, M., & Ma, X. (2023). Who should I trust: AI or myself? Leveraging human and AI correctness likelihood to promote appropriate trust in AI-assisted decision-making. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany (pp. 1–19). https://doi.org/10.1145/3544548.3581058

Madras, D., Pitassi, T., & Zemel, R. (2018). Predict responsibly: Improving fairness and accuracy by learning to defer. *Advances in Neural Information Processing Systems*, Montreal, Canada, *31*, 1–11.

Mahmud, H., Islam, A. K. M. N., Ahmed, S. I., & Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technological Forecasting & Social Change*, *175*, 1–26. https://doi.org/10.1016/j.techfore.2021.121390

Mallari, K., Inkpen, K., Johns, P., Tan, S., Ramesh, D., & Kamar, E. (2020). Do I look like a criminal? Examining how race presentation impacts human judgement of recidivism. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu, USA (pp. 1–13). https://doi.org/10.1145/3313831.3376257

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J. . . . De Fauw, J. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, *577*(7788), 89–94. https://doi.org/10.1038/s41586-019-1799-6

Mikalef, P., Conboy, K., Lundström, J. E., & Popovič, A. (2022). Thinking responsibly about responsible AI and 'the dark side' of AI. *European Journal of Information Systems*, *31*(3), 257–268. https://doi.org/10.1080/0960085X.2022.2026621

Mikalef, P., & Gupta, M. (2021). Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance. *Information & Management*, *58* (3), 1–20. https://doi.org/10.1016/j.im.2021.103434

Morrison, K., Spitzer, P., Turri, V., Feng, M., Kühl, N., & Perer, A. (2024). The impact of imperfect XAI on human-AI decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW1), 1–39. https://doi.org/10.1145/3641022

Mozannar, H., & Sontag, D. (2020). Consistent estimators for learning to defer to an expert. *Proceedings of the International Conference on Machine Learning*, *119*, 7076–7087.

Parsons, J., & Wand, Y. (2012). Extending classification principles from information modeling to other disciplines. *Journal of the Association for Information Systems*, *14*(5), 245–273. https://doi.org/10.17705/1jais.00332

Rai, A., Constantinides, P., & Sarker, S. (2019). Next generation digital platforms: Toward human-AI hybrids. *MIS Quarterly*, *43*(1), iii–ix.

Rastogi, C., Leqi, L., Holstein, K., & Heidari, H. (2023). A taxonomy of human and ML strengths in decision-making to investigate human-ML complementarity. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Delft, Netherlands (pp. 127–139). https://doi.org/10.1609/hcomp.v11i1.27554

Reverberi, C., Rigon, T., Solari, A., Hassan, C., Cherubini, P., Cherubini, A., Awadie, H., Bernhofer, S., Carballal, S., Dinis-Ribeiro, M., Fernández-Clotett, A., Esparrach, G. F., Gralnek, I., Higasa, Y., Hirabayashi, T., Hirai, T., Iwatate, M., Kawano, M. , and Tanaka, Y. (2022). Experimental evidence of effective human–AI collaboration in medical decision-making. *Scientific Reports*, *12*(1), 1–10. https://doi.org/10.1038/s41598-022-18751-2

Rinta-Kahila, T., Someh, I., Gillespie, N., Indulska, M., & Gregor, S. (2022). Algorithmic decision-making and system destructiveness: A case of automatic debt recovery. *European Journal of Information Systems*, *31*(3), 313–338. https://doi.org/10.1080/0960085X.2021.1960905

Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American*

Statistical Association, *88*(424), 1273–1283. https://doi.org/10.1080/01621459.1993.10476408

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Sanders, N. R., & Ritzman, L. P. (1991). On knowing when to switch from quantitative to judgemental forecasts. *International Journal of Operations & Production Management*, *11*(6), 27–37. https://doi.org/10.1108/01443579110005523

Sanders, N. R., & Ritzman, L. P. (1995). Bringing judgment into combination forecasts. *Journal of Operations Management*, *13*(4), 311–321. https://doi.org/10.1016/0272-6963(95)00039-9

Sanders, N. R., & Ritzman, L. P. (2001). Judgmental adjustment of statistical forecasts. In J. S. Armstrong (Ed.), *Principles of Forecasting. International Series in Operations Research & Management Science* (Vol. 30). https://doi.org/10.1007/978-0-306-47630-3_18

Sarkar, B., Shih, A., & Sadigh, D. (2023). Diverse conventions for human-AI collaboration. *Advances in Neural Information Processing Systems*, New Orleans, USA. *36*, 23115–23139

Schecter, A., Nohadani, O., & Contractor, N. (2022). A robust inference method for decision-making in networks. *Management Information Systems Quarterly*, *46*(2), 713–738. https://doi.org/10.25300/MISQ/2022/15992

Schemmer, M., Hemmer, P., Kühl, N., Benz, C., & Satzger, G. (2022). Should I follow AI-based advice? Measuring appropriate reliance in human-AI decision-making. *Workshop on Trust and Reliance in AI-Human Teams at the 2022 CHI Conference on Human Factors in Computing Systems*, New Orleans, USA (pp. 1–10).

Schmidt, C. V. H., Guffler, M., Kindermann, B., & Flatten, T. (2023). Collaborating with generative AI: Exploring algorithm appreciation in creative writing. *Proceedings of the 44th International Conference on Information Systems*, Hyderabad, India (pp. 1–9).

Schoeffer, J., Jakubik, J., Voessing, M., Kuehl, N., & Satzger, G. (2023). On the interdependence of reliance behavior and accuracy in AI-assisted decision-making. *Proceedings of the Second International Conference on Hybrid Human-Artificial Intelligence*, Munich, Germany (pp. 46–59). https://doi.org/10.3233/FAIA230074

Schunk, D. H. (1995). Self-efficacy, motivation, and performance. *Journal of Applied Sport Psychology*, *7*(2), 112–137. https://doi.org/10.1080/10413209508406961

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., Schwabe, G., & Söllner, M. (2020). Machines as teammates: A research agenda on AI in team collaboration. *Information & Management*, *57*(2), 1–22. https://doi.org/10.1016/j.im.2019.103174

Sieck, W. R., & Arkes, H. R. (2005). The recalcitrance of overconfidence and its contribution to decision aid neglect. *Journal of Behavioral Decision Making*, *18*(1), 29–53. https://doi.org/10.1002/bdm.486

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., & Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, *362*(6419), 1140–1144. https://doi.org/10.1126/science.aar6404

Simons, T., Pelled, L. H., & Smith, K. A. (1999). Making use of difference: Diversity, debate, and decision comprehensiveness in top management teams. *Academy of Management Journal*, *42*(6), 662–673. https://doi.org/10.2307/256987

Spitzer, P., Kühl, N., Heinz, D., & Satzger, G. (2023). ML-based teaching systems: A conceptual framework. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW2), 1–25. https://doi.org/10.1145/3610197

Stauder, M., & Kühl, N. (2022). AI for in-line vehicle sequence controlling: Development and evaluation of an adaptive machine learning artifact to predict sequence deviations in a mixed-model production line. *Flexible Services and Manufacturing Journal*, *34*(3), 709–747. https://doi.org/10.1007/s10696-021-09430-x

Steyvers, M., Tejeda, H., Kerrigan, G., & Smyth, P. (2022). Bayesian modeling of human–AI complementarity. *Proceedings of the National Academy of Sciences*, *119*(11), 1–7. https://doi.org/10.1073/pnas.2111547119

Swaroop, S., Buçinca, Z., Gajos, K. Z., & Doshi-Velez, F. (2024). Accuracy-time tradeoffs in AI-assisted decision making under time pressure. *Proceedings of the 29th International Conference on Intelligent User Interfaces*, Greenville, USA (pp. 138–154). https://doi.org/10.1145/3640543.3645206

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285. https://doi.org/10.1126/science.1192788

Terveen, L. G. (1995). Overview of human-computer collaboration. *Knowledge-Based Systems*, *8*(2), 67–81. https://doi.org/10.1016/0950-7051(95)98369-H

Turel, O., & Kalhan, S. (2023). Prejudiced against the machine? Implicit associations and the transience of algorithm aversion. *MIS Quarterly*, *47*(4), 1369–1394. https://doi.org/10.25300/MISQ/2022/17961

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, *8*, 2293–2303. https://doi.org/10.1038/s41562-024-02024-1

van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, *291*, 1–19. https://doi.org/10.1016/j.artint.2020.103404

Vassilakopoulou, P., Haug, A., Salvesen, L. M., & Pappas, I. O. (2023). Developing human/AI interactions for chat-based customer services: Lessons learned from the Norwegian government. *European Journal of Information Systems*, *32*(1), 10–22. https://doi.org/10.1080/0960085X.2022.2096490

Vössing, M., Kühl, N., Lind, M., & Satzger, G. (2022). Designing transparency for effective human-AI collaboration. *Information Systems Frontiers*, *24*(3), 877–895. https://doi.org/10.1007/s10796-022-10284-3

Westphal, M., Hemmer, P., Vössing, M., Schemmer, M., Vetter, S., & Satzger, G. (2024). Towards understanding AI delegation: The role of self-efficacy and visual processing ability. *ACM Transactions on Interactive Intelligent Systems*, *15*(1), 1–23. https://doi.org/10.1145/3696423

Wilder, B., Horvitz, E., & Kamar, E. (2020). Learning to complement humans. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, Yokohama, Japan (pp. 1526–1533). https://doi.org/10.24963/ijcai.2020/212

Yu, K., Berkovsky, S., Taib, R., Zhou, J., & Chen, F. (2019). Do I trust my machine teammate? An investigation from perception to decision. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, Los Angeles, USA (pp. 460–468). https://doi.org/10.1145/3301275.3302277

Zhang, Q., Lee, M. L., & Carter, S. (2022). You complete me: Human-AI teams and complementary expertise. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, New Orleans, USA (pp. 1–28). https://doi.org/10.1145/3491102.3517791

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Barcelona, Spain (pp. 295–305). https://doi.org/10.1145/3351095.3372852

Zhou, L., Paul, S., Demirkan, H., Yuan, L., Spohrer, J., Zhou, M., & Basu, J. (2021). Intelligence augmentation: Towards building human-machine symbiotic relationship. *AIS Transactions on Human-Computer Interaction*, *13*(2), 243–264. https://doi.org/10.17705/1thci.00149