Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Roles of Artificial Intelligence in Collaboration with Humans: Automation, Augmentation, and the Future of Work

Andreas Fügener, Dominik D. Walzner, Alok Gupta

informs
https://pubsonline.informs.org/journal/mnsc
MANAGEMENT SCIENCE
Vol. 72, No. 1, January 2026, pp. 538–557
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Roles of Artificial Intelligence in Collaboration with Humans: Automation, Augmentation, and the Future of Work

Andreas Fügener,[a,]* Dominik D. Walzner,[a] Alok Gupta[b]

[a] Department of Operations Management, University of Cologne, 50923 Cologne, Germany; [b] Information and Decision Sciences, Carlson School of Management, University of Minnesota, Minneapolis, Minnesota 55455

*Corresponding author

Contact: andreas.fuegener@uni-koeln.de, https://orcid.org/0000-0002-4580-7444 (AF); dominik.walzner@uni-koeln.de, https://orcid.org/0000-0002-9222-6642 (DDW); gupta037@umn.edu, https://orcid.org/0000-0002-2097-1643 (AG)

Received: April 13, 2024
Revised: December 16, 2024; May 25, 2025
Accepted: May 28, 2025
Published Online in Articles in Advance: October 15, 2025

https://doi.org/10.1287/mnsc.2024.05684

Abstract. Humans will see significant changes in the future of work as collaboration with artificial intelligence (AI) will become commonplace. This work explores the benefits of AI in the setting of judgment tasks when it replaces humans (automation) and when it works with humans (augmentation). Through an analytical modeling framework, we show that the optimal use of AI for automation or augmentation depends on different types of human-AI complementarity. Our analysis demonstrates that the use of automation increases with higher levels of between-task complementarity. In contrast, the use of augmentation increases with higher levels of within-task complementarity. We integrate both automation and augmentation roles into our task allocation framework, where an AI and humans work on a set of judgment tasks to optimize performance with a given level of available human resources. We validate our framework with an empirical study based on experimental data in which humans classify images with and without AI support. When between-task complementarity and within-task complementarity exist, we see a consistent distribution of work pattern for optimal work configurations; AI automates relatively easy tasks, AI augments humans on tasks with similar human and AI performance, and humans work without AI on relatively difficult tasks. Our work provides several contributions to theory and practice. The findings on the effects of complementarity provide a nuanced view regarding the benefits of automation and augmentation. Our task allocation framework highlights potential job designs for the future of work, especially by considering the often-ignored, critical role of human resource reallocation in improving organizational performance.

History: Accepted by D. J. Wu, Special Issue on the Human-Algorithm Connection.

Supplemental Material: The online appendix and data files are available at https://doi.org/10.1287/mnsc.2024.05684.

Keywords: artificial intelligence • human-AI collaboration • automation • augmentation • future of work

1. Introduction

The future of work will be characterized by the diffusion of artificial intelligence (AI) into almost all areas of the workplace. Indeed, recent developments in areas of deep learning (LeCun et al. 2015) have allowed AI to perform judgment tasks, such as image recognition, object detection, or speech recognition, on par with or even better than humans. Furthermore, AI can make decisions autonomously, which means that it has agentic qualities formerly exclusive to human decision makers (Baird and Maruping 2021).

In most environments that require complex decision making, human-AI collaboration can achieve even better performance than humans or AI working alone (Dellermann et al. 2019). One example is skin cancer detection. Although modern AI outperforms even experienced dermatologists (Esteva et al. 2017), studies indicate that combining AI with physician estimates can yield even better results (Hekler et al. 2019). However, a possibility that has not yet been discussed is that of a heterogeneous approach, in which physicians are sometimes substituted with AI, sometimes work together with AI, and sometimes perform cancer detection without help from AI. The human resources made available by automation in some cases could be reallocated so that groups of humans could solve more challenging cases.

In this paper, we explore AI's role in collaborating with humans on judgment tasks, assuming that humans and AI have complementary skills. We differentiate between *automation*, where AI substitutes for humans, and *augmentation*, where AI supports humans by providing advice. To our knowledge, there is a lack of structural investigation into the benefits that AI could provide through these roles, the effect of complementarity on the benefit of collaborating with AI, and the distribution of work between humans and AI.

We approach this investigation in two ways. First, we propose a task allocation framework in which tasks can be performed by one of three work constellations: by humans, by AI (automation), or by humans with AI (augmentation). Compared with the status quo setting, in which each task is performed by one human, we analyze the benefits that AI can provide through automation and augmentation. One of the primary benefits of automation is that AI can replace humans in tasks on which it performs better (*substitution benefit*). Another benefit is that the human resources released through automation can then generate value by working on other tasks (*reallocation benefit*). Through augmentation, AI provides benefit by interacting with humans and thus, improving their performance (*augmentation benefit*). We theorize that these benefits are dependent on two types of complementarity: on the potential complementarity because of performance differences between humans and AI on different tasks (*between-task complementarity*) and on the potential complementarity because of the interaction of humans and AI on the same task (*within-task complementarity*). In the second part of our investigation, we validate our findings with data from an experimental study in a context where real humans performed a set of image classification tasks either alone or with the help of AI. The setting of image classification was selected because it is a generic, idealized judgment task that is highly suitable for validating our theoretical findings. Because state-of-the-art AI solutions exist and both between-task complementarity and within-task complementarity arise, the context is suitable for validating the findings of our theoretical discussion. Our paper aims to answer the following research questions.

1. *How does the type of complementarity between humans and AI in judgment tasks affect the different benefits that AI can provide through automation and augmentation?*
2. *How does the introduction of AI for automation and augmentation in judgment tasks affect the distribution of work between humans and AI?*

Our study yields three main findings. First, we theoretically and empirically show that the benefits of AI depend on the type of complementarity. The benefits that AI can provide through automation (substitution and reallocation benefit) increase with between-task complementarity between humans and AI because humans and AI are allocated to those tasks where each can provide the greatest benefit. In contrast, the benefit that AI can provide through augmentation increases with within-task complementarity between humans and AI. Second, our framework, which combines automation, augmentation, and reallocation of the human resources released through automation, yields superior results to those obtained with only automation or augmentation. In our empirical study, humans achieved an accuracy of 68%, automation achieved 77%, and augmentation achieved 80%. With the same human resources, our framework yielded an accuracy of 88%, greatly outperforming the other options. Third, when AI is used for automation and augmentation in a setting with between-task and within-task complementarity, we observe a consistent pattern regarding the distribution of work; AI automates tasks with high relative AI performance and augments individual humans for tasks with similar AI and human performance. For tasks with low relative AI performance, humans work in crowds without AI. We argue that this pattern indicates a possible scenario for the future of work; AI takes over easy tasks, individual humans work with AI on more challenging tasks, and multiple humans work together to master the most difficult tasks.

The remainder of the paper is structured as follows. In Section 2, we provide the theoretical background for our task allocation framework and outline the contributions of this work. We justify the existence of complementarity between humans and AI, and we discuss different types of complementarity that are relevant in our framework. We also introduce automation and augmentation as potential roles that AI can perform and describe their relationships with the different types of complementarity. In Section 3, we present our task allocation framework, which allows us to analyze the effects of the two types of complementarity on the benefits that AI can provide through automation and augmentation and on the distribution of work. In Section 4, we present an empirical study in which we validate the effectiveness of our framework. The results of the empirical study are presented in Section 5. In Section 6, we elaborate on the theoretical and managerial implications of our task allocation framework, and we provide an outlook for future research.

2. Theoretical Background

In this section, we provide the theoretical background that informs our task allocation framework. Our framework rests on the assumption that complementarity between humans and AI exists. We provide extensive justification for this assumption in Section 2.1, where we discuss sources of complementarity and present different types of complementarity that are relevant for our framework. In Section 2.2, we present two roles in which AI can work alongside humans and discuss the conditions under which different types of complementarity can be realized. In the discussion of complementarity, we limit our definition of AI to supervised machine learning algorithms. Through supervised machine learning, AI infers relationships between input features and desired outputs based on historical data without having to rely on explicitly stated rules that were formulated by experts (Autor 2015). In Section 2.3, we outline the contributions of this work to the literature.

2.1. Complementarity Between Humans and AI

In contrast to the notion that the rise of AI technologies will soon lead to the replacement of human experts (Schuetz and Venkatesh 2020), several studies have shown that collaborative human-AI systems can outperform both humans and AI working alone (see Hekler et al. 2019, Tschandl et al. 2019 for examples in healthcare settings). A requirement for this performance enhancement is that complementarity between humans and AI exists, which can be leveraged in a collaborative framework (Fügener et al. 2022).

Sources of Complementarity. Complementarity between humans and AI exists because humans and AI make decisions differently. More specifically, the difference lies in the type of knowledge that humans and AI can gain and in the ways in which they use their knowledge when making judgments (Broussard 2018, Smith 2019). In the following discussion, we outline the advantages and disadvantages of humans and AI in decision-making environments.

An advantage of humans lies in their ability to make intuitive decisions. Owing to their high context awareness and ability to adapt to new situations, humans can make accurate predictions with little data (Lake et al. 2015). Humans can draw upon tacit (i.e., implicit) knowledge, which they obtain through practical action (Orlikowski 2002). However, humans have problems explicating this tacit knowledge; this is known as *Polanyi's paradox*, implying that humans cannot always construct a formalized decision process (Polanyi 1966, Autor 2014). Tacit knowledge, therefore, represents an important source of complementarity (Autor et al. 2003). A disadvantage of humans is their fallibility in judgment, which can result from their bounded rationality (Simon 1957). Because human decision makers rely heavily on heuristics, their judgment can be biased, for example, if they weight the available information ineffectively (Tversky and Kahneman 1974). Furthermore, the decisions of humans can be strongly influenced by fatigue (Blattberg and Hoch 1990) or social pressure (Asch 1955).

An advantage of AI is that it is able to analyze large amounts of data systematically because of advances in algorithmic developments combined with exponential increases in computing power. In this way, AI is able to uncover relationships in data that are difficult to find for humans (Lake et al. 2017). AI has a strong focus on formal rationality, which means that it can make decisions consistently without being affected by the same fallibility that affects human decision making (Lindebaum et al. 2020). Owing to its scalability, AI can make many decisions at negligible costs (Beam and Kohane 2016). A disadvantage of AI is that its abilities are limited to the insights that it can draw from the training data, and AI performs poorly if not enough training data are available (Agrawal et al. 2018). Even if enough training instances are available, the data may not capture all of the knowledge that is required to solve the task. Although machine learning is described by some as a way to resolve Polanyi's paradox and to close the knowledge gap between humans and AI (Brynjolfsson and Mitchell 2017, Brynjolfsson et al. 2018), Lebovitz et al. (2021) argue that the training data capture only explicit (i.e., know-what) knowledge and fail to capture the underlying uncertainty of the decision and the tacit (i.e., know-how) knowledge that experts rely on when performing tasks. Because the training data do not capture context-dependent knowledge, low crosscontext awareness can lead to a lack of generalizability (Baird and Maruping 2021, Chekroud et al. 2024).

Types of Complementarity. Given the knowledge gap between humans and AI and their distinct capabilities, we aim to capture complementarity to increase overall task performance without the need to increase the number of invested resources. We distinguish between two types of complementarity that can be realized in our task allocation framework.

Between-task complementarity is defined as complementarity that can arise because of performance differences between humans and AI across tasks (Walzner et al. 2023). Based on differences in available information and learning behavior, humans and AI can make different errors (Geirhos et al. 2021, Steyvers et al. 2022). Thus, work can be distributed so that it allows humans and AI to play to their respective strengths. Consider an example case with two tasks, A and B, where an AI performs task A with 60% accuracy and task B with 80% accuracy and a human performs task A with 80% accuracy and task B with 60% accuracy. Although the human and the AI have the same average performance of 70% over both tasks, allocating task A to the human and task B to the AI enables a combined performance of 80% (Puranam 2021).

Within-task complementarity is defined as complementarity that can arise because of the interaction of humans and AI on a task. When humans and AI work together on a task, within-task complementarity is realized if the collaborative task performance is better than the performance that either the human or AI could achieve alone on this task (Donahue et al. 2022, Walzner et al. 2023). When AI provides advice to humans, within-task complementarity can be realized if the human is able to differentiate between correct and incorrect advice, enabling augmented human-AI performance on one task to exceed both human and AI performance. Consider the example of task C, which an AI and a human both perform with 70% accuracy. If the performance of the human and the AI on task C is uncorrelated and if the human is able to differentiate between correct and incorrect advice and only follows correct advice, then advice taking could lead to an augmented performance of 91%. In 70% of the cases, the human is right. In the remaining 30%, the AI provides correct advice in 70% of the cases, yielding an improvement of 21 percentage points (Puranam 2021).

An important question is how to take advantage of the complementarity between humans and AI: that is, how to achieve a result through collaboration that is better than what humans and AI would achieve individually. This is not a trivial question as empirical results have shown that complementarity is not easily realized (Donahue et al. 2022). We analyze the impacts of utilizing AI for different types of human-AI complementarity, and we discuss how various roles of AI can help to achieve complementary performance.

2.2. Roles of AI

In our task allocation framework, we distinguish between two types of roles that AI can perform: *automation*, where AI performs tasks without human intervention, and *augmentation*, where AI performs tasks together with humans. Although many researchers on human-AI collaboration seem to regard augmentation more favorably than automation, it has been argued from a paradox-theory perspective that automation and augmentation cannot be separated as they are interdependent (Raisch and Krakowski 2021). Consistent with that logic, we analyze the effectiveness of both roles simultaneously. The type of complementarity that can be realized within our framework depends on the role that AI performs.

When AI is used for automation, it performs a task without human intervention (Raisch and Krakowski 2021). Starting from the baseline case, where each task is performed by one human, the introduction of AI for automation results in the substitution of human labor. Through automation, between-task complementarity, which can arise because of performance differences between humans and AI, can be realized. We decompose the benefit of automation into two subcomponents. The *substitution benefit* refers to the increase in performance that results from AI replacing humans. The *reallocation benefit* refers to the increase in performance that results from allocating the human who was substituted by AI to another task. Although substitution benefit has been the focus of academic as well as corporate thought processes, very little attention has been given to potential benefits that can be obtained through reallocation. In fact, automation could even be beneficial when AI performance is inferior if the reallocation benefit exceeds the negative substitution benefit. This could especially be important in situations where human resources are scarce.

When AI is used for augmentation, it collaborates with humans to generate a combined judgment (Jain et al. 2021). Although many augmentation mechanisms exist, we focus on advice taking in our analytical framework as it is one of the most prominent forms of decision augmentation (Bansal et al. 2019). In such a case, AI is not directly involved in working on a task but only in providing decision support to a human decision maker (Maedche et al. 2019). The collaborative setting in which an AI provides advice to a human can be modeled with the so-called judge-advisor system (Bonaccio and Dalal 2006). In the context of human-AI collaboration, AI takes the role of the advisor, whereas the human takes the role of the judge who is free to consider or dismiss the advice (Bansal et al. 2019). Through augmentation, within-task complementarity, which can arise through the interaction of humans and AI, can be realized. The *augmentation benefit* refers to the increase in performance that results from the AI providing advice to the human. The realization of the augmentation benefit depends on the entity that is responsible for combining the judgments of humans and AI. For advice taking, it is the responsibility of humans to integrate AI advice into their final decisions. The effectiveness of augmentation, therefore, rests on the ability of humans to distinguish good advice from bad advice: that is, to learn a good mental model for determining when AI is correct (Bansal et al. 2019) and for the task itself (Jussupow et al. 2021).

2.3. Contribution to the Literature

Our research contributes to two streams of literature. The first literature stream explores effective collaboration designs for humans and AI that can be viewed as constituting a "multiagent, goal-oriented system" (Puranam 2021, p. 76). In this system, different work constellations can be considered for a task where AI works alone (automation), works together with humans (augmentation), or is not used at all. Given the different types of complementarity discussed above, a crucial question for organizations is when to prioritize which work constellation. Empirical evidence supports the idea that organizations lose potential when they rely only on a single work constellation. Kesavan and Kushwaha (2020) conducted a field experiment at a spare automobile parts retailer, where humans had the option to override a data-driven decision-making tool. They found that humans overriding the tool (augmentation) performed better than the tool (automation) for growth-stage products, for which little data are available. In contrast, augmentation performed worse than automation for mature- and declining-stage products, for which large amounts of data are available. Their results indicate that tasks with little data are difficult for AI tools, which means that these are tasks for which humans can provide complementary knowledge. In contrast, tasks with large amounts of data may be relatively easy for AI tools.

To determine which work constellation to use, different frameworks have been proposed in the literature. So-called triage models have been introduced as a way to decide ex ante who should perform a given (prediction) task. A triage model is a type of metamodel that is trained to predict when an intervention by a human is beneficial for task performance (Raghu et al. 2019). Many recent papers discuss the learning-to-defer approach, where the AI learns to decide whether it should perform a task on its own or should defer the task to a human (Madras et al. 2018, Mozannar and Sontag 2020). Chen et al. (2023) presented a framework to decide when to rely on human adjustments to algorithmic demand forecasts (augmentation) and when to rely only on the algorithmic forecast (automation). Although all of these frameworks provide helpful insights into effective human-AI collaboration design, structural investigation regarding the factors that drive the effectiveness of each work constellation is lacking. Furthermore, many of these frameworks neglect the possibility of tasks being performed by humans without the involvement of AI, which is particularly relevant if humans work in groups or crowds. We contribute to this literature stream by designing a novel framework in which a task can be allocated to humans, AI (automation), or both humans and AI (augmentation). Based on this framework, our main contributions are defining the benefits that AI can provide through automation and augmentation and providing a theoretical discussion of how the different types of complementarities affect the benefits of AI. The benefit of automation (which can be decomposed into the substitution and reallocation of humans) depends on the level of between-task complementarity, whereas the benefit of augmentation depends on the level of within-task complementarity. In our empirical analysis, we show that our task allocation framework leads to significant performance improvements compared with settings in which all tasks are performed by humans, automation, or augmentation.

The second literature stream studies the future of work with a specific focus on integrating AI into the workplace. Because jobs can be viewed as a set of tasks (Autor et al. 2003), it is theorized that the impact of AI needs to be analyzed at the task level as it will lead to more complex outcomes than the simple replacement of humans (Autor 2015, Brynjolfsson and Mitchell 2017). Based on the complementary skills of humans and AI discussed above, different work constellations will emerge, in which AI will be used to automate some tasks and augment humans in other tasks (Rai et al. 2019, Puranam 2021). Some tasks will continue to be performed by humans without AI because AI cannot (yet) provide sufficient value (Acemoglu and Restrepo 2018, Puranam 2021). Existing research on the impacts of AI on the future of work has mainly examined different AI roles in isolation, focusing on the impact of either automation or augmentation. For example, Acemoglu and Restrepo (2018) developed a model that captures the impact of automation on human labor. In this model, the automation of existing tasks leads to the creation of new tasks in which humans have superior skills. To obtain a more complete view of the impact of AI, one needs to look at both automation and augmentation together because these roles exhibit strong interdependencies (Raisch and Krakowski 2021). Difficulties in predicting the combined impact of automation and augmentation arise from shortages of task-level models that consider properties such as the complementarity between humans and AI (Frank et al. 2019). Recent work by Dell'Acqua et al. (2023, p. 1) suggests that the abilities of AI lie along a "jagged technological frontier" and that this frontier determines the benefits that AI can provide. Based on our task allocation framework, we theoretically analyze when a task should be performed by humans, automation, or augmentation. In our analysis, the dominance of each work constellation depends on the level of between-task and within-task complementarity and on the difficulty of the task for the AI. The two types of complementarity and the task difficulty from the AI's perspective help conceptualize this "jagged technological frontier of AI capabilities." Our empirical results validate the theoretical discussion of the distribution of work; AI automates easy tasks, AI augments humans on medium tasks, and humans work in groups without AI on difficult tasks. Based on further empirical analyses, we provide an indication about how the distribution of work will be affected by significant improvements in AI capabilities.

3. Task Allocation Framework

In this section, we present a novel task allocation framework that allocates humans and AI to a given set of judgment tasks. The framework is inspired by Michelucci and Dickinson (2016), who present a conceptual overview of distributed systems in a crowdsourcing context. We use our framework to analyze the distribution of work between humans and AI as well as the benefits that AI can provide through automation and augmentation. Here, benefit is defined as the improvement in task performance relative to the baseline case, in which each task is performed by one human. Figure 1 illustrates the framework.

To describe the general characteristics of the type of tasks that we deem appropriate for our framework, we focus on two conditions. First, the *Polanyi condition* states that the procedure for optimally solving the tasks cannot be formalized; therefore, these tasks cannot be solved—either by AI or by humans—with absolute certainty (Autor 2015). Thus, a human or an AI will solve those tasks differently. This condition ensures the existence of complementarity, which is necessary to generate benefits through human-AI collaboration. An example of a task that does not satisfy this condition is a deterministic optimization problem, for which the optimal solution can simply be calculated. Here, no complementary knowledge of humans and AI exists. Second, the *ground truth condition* requires that the tasks follow a certain structure. More specifically, the task needs to have one true answer (ground truth), which the decision maker must ascertain from a set of options (McGrath 1984). Thus, humans and AI can solve the task correctly with a certain probability. The condition also states that a meaningful aggregation mechanism for crowd-based decision making exists, which enables the combination of individual choices into an aggregated choice: for example, through majority voting for classification tasks or through averaging for regression tasks.

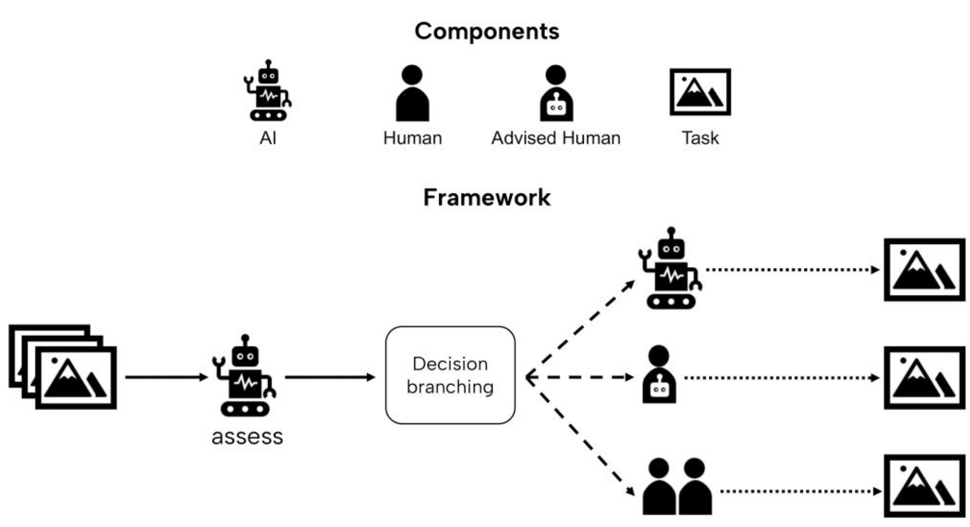
3.1. Basic Notation

We consider a set of classification tasks $t \in \mathcal{T} = \{1, \ldots, T\}$, where T denotes the number of tasks. For each task, the correct choice must be selected from a set of available options $c \in \mathcal{C} = \{1, \ldots, C\}$, where C denotes the number of options. In our framework, tasks can be allocated to one or more humans, AI (automation), or one or more humans with AI (augmentation). Whenever a task is allocated to more than one human, we assume a wisdom-of-crowds setting, in which the aggregated decision is determined based on majority voting (Boland et al. 1989). We formally define the performance of each work constellation below.

Humans. We denote p_{tc}^H as the probability of a human choosing option c for task t, where $\sum_{c \in \mathcal{C}} p_{tc}^H = 1$. Let p_t^H denote the probability of a human choosing the correct option, and let the expected performance of a human on a set of tasks $\mathcal{T}$ be $1/T \sum_{t \in \mathcal{T}} p_t^H$. Note that we assume heterogeneous performance by humans, but the framework does not know which human will work on a task. Thus, p_t^H can be interpreted as the expected performance or the average performance of a randomly selected human on task t.

When multiple humans work together, we denote the performance of n humans working on task t as $P_{nt}^H(p_t^H)$. If a task is allocated to one human, then $P_{1t}^H(p_t^H) = p_t^H$. If a task is allocated to multiple humans, we apply a majority voting scheme, and the performance equals the probability that the humans select the correct choice over any other choice. The mathematical characteristics of majority voting have been discussed by Condorcet (1785) for the case of two options and by List and Goodin (2001) for the case of more than two options. Assuming independence between individual crowd members, they show that if the probability of

Figure 1. Task Allocation Framework



choosing the correct option is higher than all of the other options, then it is more likely to be chosen as the crowd option than the other options. Further, they also show that with an increasing number of crowd members, the probability of the crowd choosing the correct option converges to one.

AI (Automation). When AI is used for automation, it performs a task by allocating a likelihood score to each option, which represents the certainty of the AI regarding whether the option is correct. It then selects the option with the highest score. We use the likelihood score l_t^{AI} as an estimator of the probability p_t^{AI} that the AI performs the task correctly.

Humans with AI (Augmentation). When AI is used for augmentation, it recommends an option to a human decision maker, who can then decide whether to follow or ignore it. The probability of following the AI advice is denoted as the weight of advice r. We define the weight of advice as the relative increase in the probability of selecting choice c if it is recommended by the AI. A potential benefit of augmentation in the case of advice taking is to assume that incorrect advice may be discounted compared with correct advice. This assumption is consistent with the advice-taking literature, which shows that humans are (to some extent) able to differentiate between correct and incorrect advice (Yaniv and Kleinberger 2000, Bonaccio and Dalal 2006). We model this by differentiating the weight of correct advice r^{AI} and the weight of incorrect advice $r^{\overline{AI}}$, which are both defined in [0, 1]. Thus, the performance of a human who receives correct advice increases to $p_t^H + r^{AI}(1 - p_t^H)$, whereas the performance of a human who receives incorrect advice decreases to $p_t^H - r^{\overline{AI}} p_t^H$. As the probabilities of correct and incorrect advice are p_t^{AI} and $(1 - p_t^{AI})$, respectively, the performance of multiple humans working with AI advice can be calculated as follows:

$$P_{nt}^{HAI}(p_t^H, p_t^{AI}, r^{AI}, r^{\overline{AI}}) = p_t^{AI} \cdot P_{nt}^H(p_t^H + r^{AI}(1 - p_t^H)) + (1 - p_t^{AI}) \cdot P_{nt}(p_t^H - r^{\overline{AI}} p_t^H). \quad (1)$$

For the special case of one human working with AI, the performance is

$$P_{1t}^{HAI}(p_t^H, p_t^{AI}, r^{AI}, r^{\overline{AI}}) = p_t^{AI}(p_t^H + r^{AI}(1 - p_t^H)) + (1 - p_t^{AI})(p_t^H - r^{\overline{AI}} p_t^H). \quad (2)$$

This formulation is an extension of that of Boland et al. (1989), who provided a model for wisdom-of-crowds decision making with dependent decision makers. Our formulation adapts their model to the case of advice taking and extends the model by considering the heterogeneous effects of correct and incorrect AI advice. The effect of augmenting a crowd of humans with AI depends on whether the correct choice is the most likely choice before and after receiving advice. In the first case, humans without advice select the correct choice with the highest probability, and their performance converges to one as the crowd size increases. If the correct choice remains the most likely choice after receiving incorrect advice, the performance converges to one as well. If an incorrect choice is the most likely choice after receiving incorrect advice, the performance converges to that of AI. In this case, advice is detrimental for a large number of humans, and providing AI advice decreases performance by $1 - p_t^{AI}$. In the second case, an incorrect choice is the most likely choice without advice, and the performance converges to zero. If the incorrect choice remains the most likely choice after receiving correct advice, the performance converges to zero as well. If the correct choice is the most likely choice after receiving correct advice, the performance converges to the performance of AI. In this case, providing advice is beneficial for a large number of humans but is limited to the benefit of automation as providing AI advice increases performance by p_t^{AI}. Note that we refrain from discussing cases in which the probabilities of the correct choice and the most likely incorrect choice are exactly equal before and after receiving advice.

3.2. Optimization Model

We propose the task allocation framework as a simple optimization problem with the objective of maximizing average task performance over a fixed set of tasks with a fixed amount of human resources. The model allocates each task to one or more humans working alone, AI working alone (automation), or one or more humans working with AI (augmentation). The model provides a tactical allocation of humans and AI to tasks without focusing on which specific human works on which specific task. Thus, we ignore restrictions on the number of tasks that a single human can complete but instead, rely on distribution-of-effort modeling (Karush 1962), where the set of tasks competes for a fixed amount of human resources. The idea is that although one unit of human resources is sufficient to complete a task, the probability of correctly solving a task increases with the number of human resources allocated to the task.

We make the following assumptions. Each task consumes the same amount of human resources if it is allocated to a human. A sufficient number of individual human workers exist such that multiple humans can be allocated to a task. Thus, when a task is allocated to multiple units of human resources, we assume that these resources come from different humans, which is a requirement for our wisdom-of-crowds logic. The total capacity of human resources is R task units; that is, each unit of human resources relates to the completion of one task. For example, consider a case in which 1 hour of human work is required to complete a task and in total, 100 hours of human resources are available. If 100 tasks need to be completed, each task could be allocated to one human, or 80 tasks could be allocated to the AI, with the remaining 20 tasks allocated to five humans each. In this example, it would not matter whether the 100 hours of human resources are provided by 100 humans with 1 work hour each or 5 humans with 20 work hours each.

Indices

- $t \in \mathcal{T} = \{1, \ldots, T\}$: tasks
- $n = 1, \ldots N$: number of humans allocated to a task

Decision Variables

- $x^H_{nt} \in \{0,1\} \quad \forall t \in \mathcal{T},\ n = 1, \ldots N$: n humans work on task t
- $x^{AI}_t \in \{0,1\} \quad \forall t \in \mathcal{T}$: AI works on task t (automation)
- $x^{HAI}_{nt} \in \{0,1\} \quad \forall t \in \mathcal{T},\ n = 1, \ldots N$: n humans work with AI on task t (augmentation)

Parameters

- $p^H_t \in (0,1) \quad \forall t \in \mathcal{T}$: probability of one human selecting the correct choice for task t
- $P^H_{nt} \in (0,1) \quad \forall t \in \mathcal{T},\ n = 1, \ldots N$: probability of n humans selecting the correct choice for task t
- $p^{AI}_t \in (0,1) \quad \forall t \in \mathcal{T}$: probability of the AI selecting the correct choice for task t
- $P^{HAI}_{nt} \in (0,1) \quad \forall t \in \mathcal{T},\ n = 1, \ldots N$: probability of n humans with AI selecting the correct choice for task t
- $r^{AI} \in (0,1)$: weight of correct advice
- $r^{\overline{AI}} \in (0,1)$: weight of incorrect advice
- R: amount of available human resources (in task units)

Objective Function

$$\text{Max } \frac{1}{T}\sum_{t=1}^{T}\left(\sum_{n=1}^{N} x^H_{nt} \cdot P^H_{nt}(p^H_t) + x^{AI}_t \cdot p^{AI}_t + \sum_{n=1}^{N} x^{HAI}_{nt} \cdot P^{HAI}_{nt}(p^H_t, p^{AI}_t, r^{AI}, r^{\overline{AI}})\right). \tag{3}$$

Constraints

$$\sum_{n=1}^{N} x^H_{nt} + x^{AI}_t + \sum_{n=1}^{N} x^{HAI}_{nt} = 1, \qquad \forall t \in \mathcal{T} \tag{4}$$

$$\sum\nolimits_{t=1}^{T}\left(\sum_{n=1}^{N} x^H_{nt} \cdot n + \sum_{n=1}^{N} x^{HAI}_{nt} \cdot n\right) \leq R. \tag{5}$$

Objective function (3) maximizes the average performance over all tasks. Constraints (4) require that each task is performed by exactly one work constellation (one or more humans working alone, AI working alone, or one or more humans working with AI), and Constraint (5) ensures that the human resources allocated to tasks do not exceed the available human resources. The idea of the framework is straightforward; the algorithm decides whether humans or the AI should complete the task. When humans are allocated to a task, the algorithm also determines whether they should be augmented with AI advice. The mechanisms behind this valuation may be complex. For example, augmenting humans on one task might yield benefits if a small number of humans are allocated to this task; however, it might be detrimental if more humans are added as discussed in the previous section. In the following, we structurally investigate in which cases humans working without AI, automation, or augmentation provide the greatest benefit.

3.3. The Impact of Complementarity

In this section, we formalize the effect of between-task and within-task complementarity on the benefits that AI can provide through automation and augmentation. *Between-task complementarity* describes the performance difference of humans and AI between tasks. To measure between-task complementarity, we consider a simple linear model of the performance of humans and AI for a set of tasks:

$$p^H_t = a + b \cdot p^{AI}_t, \quad \forall t \in T. \tag{6}$$

Here, b serves as measure of between-task complementarity, where larger values of b indicate lower levels of between-task complementarity. Please note that this measure of between-task complementarity between humans and AI can be determined for any set of (judgment) tasks for which performance of humans and AI can be estimated. For example, we could determine the between-task complementarity for a homogenous set of image classification tasks but also, for a heterogenous set of tasks consisting of image classification and stock-out prediction tasks. In the case of $b = -1$, humans perform best (worst) whenever AI performs worst (best), $b = 0$ indicates uncorrelated human and AI performance, and $b = 1$ indicates that humans perform best (worst) whenever AI performs best (worst). *Within-task complementarity* describes the potential benefit obtained through the interaction of humans and AI on a task. As shown in Equation (1), the performance of augmented humans increases with the weight of correct advice r^{AI} and decreases with the weight of incorrect advice $r^{\overline{AI}}$. Thus, small values of r^{AI} and large values of $r^{\overline{AI}}$ indicate low levels of within-task complementarity, whereas large values of r^{AI} and small values of $r^{\overline{AI}}$ indicate high levels of within-task complementarity. We define the relative weight of advice as the ratio of the weights of correct and incorrect advice as a measure of within-task complementarity: $\hat{r} = r^{AI}/r^{\overline{AI}}$. If $\hat{r} > 1$, then humans place more weight on correct advice than on incorrect advice. If $\hat{r} \leq 1$, then humans place at least as much weight on incorrect advice as on correct advice.

3.4. Benefits of AI Collaboration

As described in the previous sections, the benefits that AI provides through automation and augmentation are driven by different effects: that is, the substitution and reallocation benefits because of automation and the augmentation benefit. When formalizing the benefits of AI, we consider every task being performed by one human without AI as the baseline case.

Through automation, AI can provide benefit by substituting for humans in tasks where it has better performance. This *substitution benefit* is calculated as the difference between AI performance and human performance for all tasks that are allocated to AI:

$$V^{sub} = \frac{1}{T}\sum_{t=1}^{T} x_t^{AI} \cdot (p_t^{AI} - p_t^H). \tag{7}$$

Furthermore, automation can provide additional benefit because humans substituted by AI can be reallocated to other tasks. This *reallocation benefit* is calculated as the difference between crowd performance and individual human performance for all tasks that are allocated to humans (alone or augmented):

$$V^{re} = \frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{N} (x_{nt}^H + x_{nt}^{HAI}) \cdot (P_{nt}^H(p_t^H) - p_t^H). \tag{8}$$

Through augmentation, AI can provide benefit by providing advice to humans. This *augmentation benefit* is calculated as the difference between the performance of humans working with AI and that of humans working without AI for all tasks that are allocated to augmented humans:

$$V^{aug} = \frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{N} x_{nt}^{HAI} \cdot (P_{nt}^{HAI}(p_t^H, p_t^{AI}, r^{AI}, \overline{r^{AI}},) - P_{nt}^H(p_t^H)). \tag{9}$$

The sum of the substitution, reallocation, and augmentation benefits constitutes the improvement in performance that our framework achieves compared with having one human working alone on each task.

3.5. Distribution of Work

For different levels of between-task and within-task complementarity, we can define for a single task and a single human resource whether the task should be allocated to humans, automation, or augmentation. All of the proofs for this section are provided in Online Appendix A. Because we only consider the distribution of work for a single task, no reallocation benefit can be obtained. In Online Appendix B, we discuss how the reallocation benefit affects the distribution of work. For notational convenience, we define the relative AI performance $\widehat{p_t^{AI}}$, where $\widehat{p_t^{AI}} < 1$ ($\widehat{p_t^{AI}} > 1$) refers to lower (higher) expected AI performance compared with expected human performance:

$$\widehat{p_t^{AI}} = \frac{\frac{p_t^{AI}}{1-p_t^{AI}}}{\frac{p_t^H}{1-p_t^H}}. \tag{10}$$

Automation Better Than Augmentation. Automation achieves better task performance than augmentation if the relative AI performance is greater than the ratio of ignoring incorrect advice and ignoring correct advice. Thus, automation provides greater benefits than augmentation for larger values of relative AI performance, whereas it provides smaller benefits than augmentation for larger weights of correct advice and smaller weights of incorrect advice:

$$\begin{aligned} &p_t^{AI} > p_t^{AI}(p_t^H + r^{AI}(1-p_t^H)) + (1-p_t^{AI})(p_t^H - \overline{r^{AI}} p_t^H) \\ &\Longleftrightarrow \widehat{p_t^{AI}} > \frac{1-\overline{r^{AI}}}{1-r^{AI}}. \end{aligned} \tag{11}$$

Automation Better Than Humans. Automation achieves better task performance than humans if the relative AI performance is greater than one. Thus, automation provides greater benefits than humans for larger values of relative AI performance:

$$p_t^{AI} > p_t^H \Longleftrightarrow \widehat{p_t^{AI}} > 1. \tag{12}$$

Augmentation Better Than Humans. Finally, augmentation achieves better task performance than humans if the relative AI performance is greater than the ratio of incorrect and correct advice. Thus, augmentation provides greater benefit than humans for larger values of relative AI performance, whereas it provides smaller benefit than humans for larger weights of incorrect advice and smaller weights of correct advice:

$$\begin{aligned} &p_t^{AI}(p_t^H + r^{AI}(1-p_t^H)) + (1-p_t^{AI})(p_t^H - \overline{r^{AI}} p_t^H) > p_t^H \\ &\Longleftrightarrow \widehat{p_t^{AI}} > \frac{\overline{r^{AI}}}{r^{AI}}. \end{aligned} \tag{13}$$

Dominant Work Constellation. With the information above, we can determine which work constellation is dominant depending on between-task and within-task complementarity. To reduce the number of measures for within-task complementarity and enable better visualization, we normalize the weights of advice so that $r^{AI} + \overline{r^{AI}} = 1$. Thus, the relative weight of correct advice $\hat{r}$ reduces to

$$\hat{r} = \frac{r^{AI}}{\overline{r^{AI}}} = \frac{r^{AI}}{1-r^{AI}}. \tag{14}$$

As the right-hand side of Equation (11) simplifies to $(1-\overline{r^{AI}})/(1-r^{AI}) = r^{AI}/\overline{r^{AI}} = \hat{r}$ and that of Equation (13) simplifies to $\overline{r^{AI}}/r^{AI} = 1/\hat{r}$, all of the indifference curves between automation, augmentation, and humans depend only on the relative weight of correct advice and the relative AI performance. As the relative AI performance is based on AI and human performance and as the estimation of human performance depends on between-task complementarity, we can determine the best-performing mode based on AI performance, between-task complementarity (measured by parameter b), and within-task complementarity (measured by parameter $\hat{r}$).

In Figure 2, we illustrate the selected work constellation based on AI performance (horizontal axes) and within-task complementarity: that is, the relative weight of correct advice (vertical axes). In the left panel of Figure 2, we assume a low level of between-task complementarity with $p_t^H = 0.1 + 0.8 \cdot p_t^{AI}$, and in the right panel of Figure 2, we assume a high level of between-task complementarity with $p_t^H = 0.4 + 0.2 \cdot p_t^{AI}$. Please note that in both cases, the average human and AI performances over all tasks are equal. Augmentation may be preferable when the relative weight of correct advice is greater than one as otherwise, the augmented performance cannot exceed the performance of automation or humans. For a given level of relative weight of correct advice, augmentation is most promising in cases where human performance and AI performance are similar. Human performance and AI performance are more often similar if the between-task complementarity is low. Thus, augmentation is more preferable for lower levels of between-task complementarity. Based on the theoretical discussion provided in this section, we formulate the following statements.

Statement 1. For a constant performance difference between humans and AI, higher levels of between-task complementarity lead to

a. a higher share of substitution benefit,
b. a higher share of reallocation benefit, and
c. a lower share of augmentation benefit.

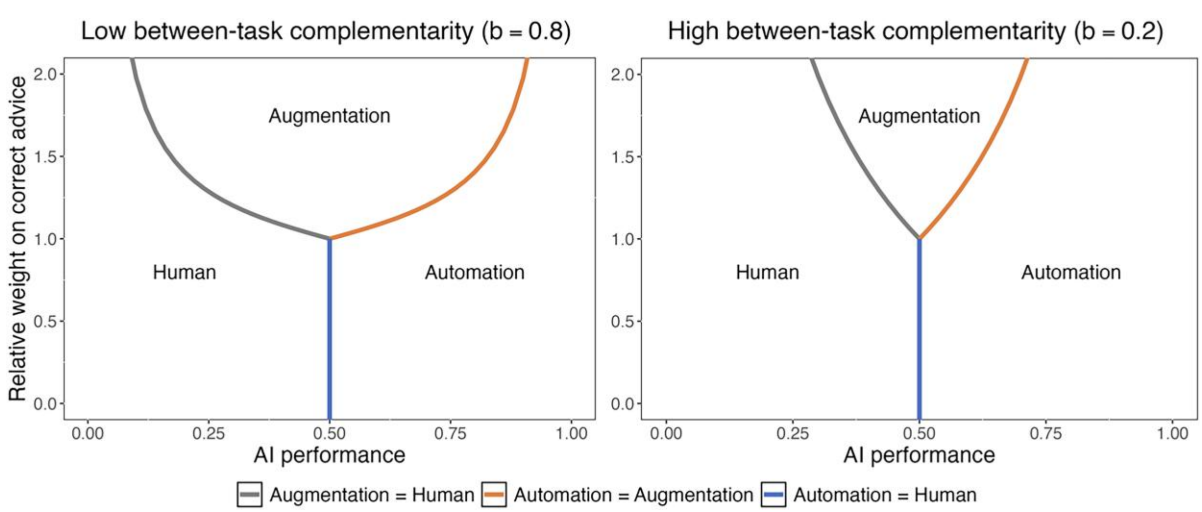
Statement 2. For a constant performance difference between humans and AI, higher levels of within-task complementarity lead to

a. a lower share of substitution benefit,
b. a lower share of reallocation benefit, and
c. a higher share of augmentation benefit.

4. Empirical Study

In this section, we present our empirical study based on experimental data to validate our theoretical discussion regarding the benefits of automation and augmentation and the distribution of work. We chose the context of image classification because of its generic and idealized nature, which makes it well suited for validating our theoretical findings. This context features clearly defined tasks, AI algorithms with performance comparable with or even better than that of humans, and complementarity both between and within tasks. Image classification also satisfies the two conditions that we defined in the previous section. The *Polanyi* condition is met because image classification is a task in which the human decision-making process cannot be completely formalized. Thus, there is a profound difference in how humans and AI classify images (Geirhos et al. 2020). Humans classify images based on tacit knowledge grounded in their experiences and do not require a lot of training for this task (Lake et al. 2015). Although image classification used to be a task in which algorithms had inherent limitations (Minsky and Papert 1969), modern AI achieves performance comparable with or even better than that of humans through deep learning technologies (Russakovsky et al. 2015, Szegedy

Figure 2. (Color online) Dominant Work Constellation Depending on AI Performance, Within-Task Complementarity (Relative Weight on Correct Advice), and Between-Task Complementarity



et al. 2016). Empirical evidence suggests that humans and AI make different kinds of errors in image classification tasks (Rosenfeld et al. 2018, Geirhos et al. 2020), which is an indication of between-task complementarity between humans and AI. We provide a discussion about what future developments might affect the complementarity between humans and AI in Section 6. Furthermore, the *ground truth* condition is also met because a ground truth exists for each image that is used in the experiment. Additionally, the choices of individual humans can be combined into an aggregated choice through majority voting.

4.1. Experimental Data (Fügener et al. 2021)

We base our empirical study on data from an experiment that was originally published by Fügener et al. (2021). We limit ourselves here to a high-level description of the experiment; more details on the experimental procedure as well as treatment screenshots can be found in Online Appendix C and in Fügener et al. (2021). In the experiment, humans had to classify a set of 100 images sourced from the ImageNet database (https://www.image-net.org/). The experiment comprised three treatments. In the first treatment, humans classified the images without help from AI. In the second and third treatments, humans received advice in the form of a class recommendation from AI: that is, the GoogLeNet Inception v3 algorithm (Szegedy et al. 2016). For each image, the AI calculated a likelihood score for each of the possible classes. This score represents the probability that a particular class is the correct class for the given image and thus, indicates the expected AI performance for this task. This likelihood score also informs on the task difficulty from the AI's perspective. Among the 100 images, the AI classified 77 correctly. In the third treatment, humans additionally received information on the AI's certainty for the recommended class. For our empirical study, we use the data from the first two treatments, where the data of humans from treatment 1 are used to simulate the performance of humans, whereas the data of humans from treatment 2 are used to simulate the performance of humans with AI (augmentation). The correctness of the class recommendations from the GoogLeNet Inception v3 algorithm represents the ex post performance of AI (automation), whereas the likelihood scores for the recommended classes l_t^{AI} represent the ex ante estimation of the AI's expected performance. In Online Appendix F, we perform a robustness check based on the humans from treatment 1 and treatment 3, replicating all of the main results.

4.2. Simulation and Evaluation

We simulate the expected performance of humans by sampling from treatment 1, and we simulate the expected performance of augmentation by sampling from treatment 2. Please note that we treat the results from these treatments as populations in our simulation study. Using Monte Carlo simulation with 10,000 iterations, we create crowds with sizes n varying from 1 to 10 by sampling n humans from the respective treatment group. If $n > 1$, we use majority voting to determine the aggregated choice. If two or more choices receive the same number of votes, then one of these choices is selected randomly.

Based on the AI's likelihood score and the simulated performance of humans and augmentation, we conduct two sets of linear regressions. The first set of regressions captures the relationship between the AI's expected performance l^{AI} and the performance of humans P^H on the set of images that were used in the experiment. The second set of regressions captures the relationship between the AI's expected performance l^{AI} and the performance of augmentation P^{HAI}. We create linear regressions for each potential number of humans allocated to a task n so that we can estimate the performance for different crowd sizes of humans and of humans working with AI based on the ex ante estimated AI performance. Thus, although our framework knows about general human performance and about between-task and within-task complementarity at an aggregated level, it has no specific information about individual images or individual humans. We provide the linear regression equations in Online Appendix D. We use the expected performance estimates of humans and augmentation based on the AI's expected performance as inputs for our task allocation framework. We determine the optimal task allocation based on Equation (3) under the assumption that we have enough individual human workers so that up to 10 humans could be allocated to any task. Based on the optimal task allocation, we then simulate the ex post performance of our framework with 10,000 simulation runs. If a task has been allocated to AI (automation), we use the realized AI performance achieved for that task. If a task has been allocated to n humans working alone, we sample n humans from treatment 1 and determine their accuracy. If a task has been allocated to n humans working with AI (augmentation), we sample n humans from treatment 2 and determine their accuracy. Thus, although the tasks are allocated based on the ex ante information about expected performance, we report the realized performance of our framework when discussing the benefits of AI. We also determine the performance of the full-automation and full-augmentation benchmarks. Furthermore, we decompose the benefit of our framework into substitution, augmentation, and reallocation benefit based on Equations (7)–(9). The results of this analysis are shown in Section 5.2 as well as in Online Appendix E.

We replicate the results of our empirical study in a robustness check where we use humans from treatment

3 to simulate the performance of humans with AI (augmentation). In addition to the AI advice, the humans in treatment 3 received information on AI certainty. The robustness check allows us to investigate whether information on the expected quality of AI advice affects within-task complementarity because of potential changes in the relative weight of correct advice. We show that although providing AI certainty does (marginally) affect within-task complementarity, the benefits from collaboration and the resulting task allocation structure remain largely unaffected. All details of the robustness check are presented in Online Appendix F.

5. Results

Before analyzing the benefits of AI and the resulting distribution of work, we establish the existence of between-task and within-task complementarity between humans and AI.

For between-task complementarity, we analyze the relationship between human performance and AI performance for the 100 images in our data set. Figure 3 shows the AI certainty (as the ex ante performance indicator) and average performance of humans for each image. Using Equation (6) for the relationship between human performance and AI performance yields the following approximation: $p_t^H = 0.44 + 0.31 p_t^{AI}$. The small value (considerably below a value of one) for the regression slope b indicates that humans and AI classify the images differently. Thus, we confirm the existence of between-task complementarity between humans and AI.

For within-task complementarity, we compare the performance of humans working without AI, that of automation, and that of augmentation for all 100 images. Humans achieved an accuracy of 68%, and AI alone (automation) achieved an accuracy of 77%. When humans and AI worked together (augmentation), they achieved an accuracy of 80%, better performance than either humans or AI achieved alone. When analyzing the weights that people place on average on correct and incorrect advice, we find that the average weight of correct advice is 0.78 and that the average weight of incorrect advice is 0.52. This results in a relative weight of correct advice of 1.52. Thus, we confirm the existence of within-task complementarity as humans are able to differentiate between correct and incorrect AI advice.

5.1. Benefits of Automation and Augmentation

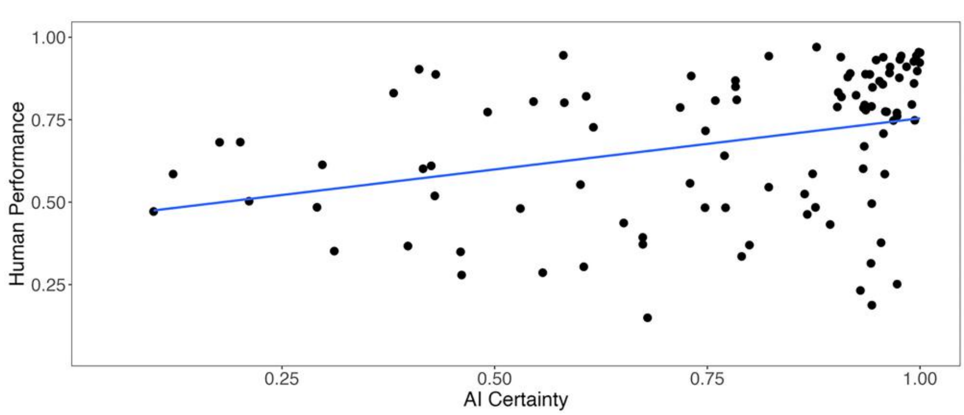
We calculate the performance benefits of the task allocation framework in comparison with the baseline case, where each image is classified by one human. To gain a better understanding of the benefits of automation and augmentation, we decompose the total benefit into substitution benefit, augmentation benefit, and reallocation benefit. We report the results of this and the subsequent analysis in tabular form in Online Appendix E, where we also report the standard errors of our empirical simulation. The low standard errors show that our results are robust with respect to sampling variations. All reported mean differences are significant (Wilcoxon test, $p < 0.01$).

Figure 4 illustrates the benefits that AI can provide through automation and augmentation. With the same level of human resources as in the baseline case, our framework achieves performance improvements of 20 percentage points: that is, an increase from an accuracy of 68% to an accuracy of 88%. These benefit gains strongly exceed those of full automation (77%) and full augmentation (80%). We observe that the benefits of AI arise entirely through substitution and reallocation, whereas augmentation does not provide any benefit.

5.2. Effects of Complementarity on Automation and Augmentation

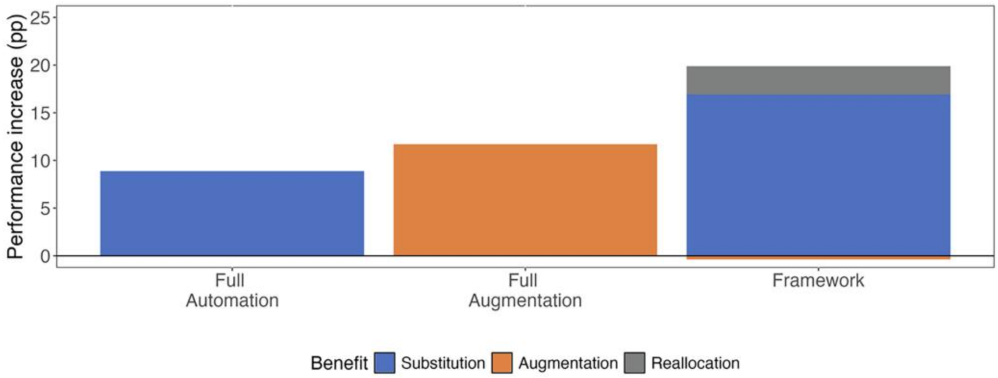
In Section 3, we formalize the effects of between-task and within-task complementarity on the benefits of automation and augmentation. To validate our theoretical formalization, we apply three different methods of

Figure 3. (Color online) AI Certainty and Human Performance per Image



Note. Each dot represents one image.

Figure 4. (Color online) Benefits of Automation and Augmentation



Note. pp, percentage points.

splitting the 100 images into two subsets of 50 images each and repeat the analysis, where we compare the benefits of the framework with those of the full-automation and full-augmentation benchmarks. First, we split the images into one group with low between-task complementarity and one group with high between-task complementarity, ensuring that both groups have similar absolute differences in average expected human performance $\overline{p^H}$ and average expected AI performance $\overline{p^{AI}}$. Between-task complementarity is expressed by the measure b. Second, we split the images into one group with low within-task complementarity and one group with high within-task complementarity, again with similar absolute differences in $\overline{p^H}$ and $\overline{p^{AI}}$. Within-task complementarity is expressed by the measures r^{AI} and $r^{\overline{AI}}$. Third, we split the images into one group with inferior AI performance and one group with superior AI performance. Summary statistics about the groups are shown in Table 1, where we also include the statistics of the full data set of 100 images as a reference. The regression coefficients required for the simulation are shown in Online Appendix D.

In Figure 5, we show the performance benefits of our framework for images with low between-task complementarity (left panel) and images with high between-task complementarity (right panel). The first decomposition validates Statement 1. Higher between-task complementarity is associated with a greater share of benefits from automation (substitution and reallocation), and lower between-task complementarity is associated with a greater share of benefit from augmentation. In addition to consistency with our first statement, we observe much greater benefits for the group with high between-task complementarity. Note that in the case of low between-task complementarity, the benefits of the framework do not outperform the benefit of full automation. Thus, in cases with low between-task complementarity and better-performing AI, it is difficult for humans to create additional value.

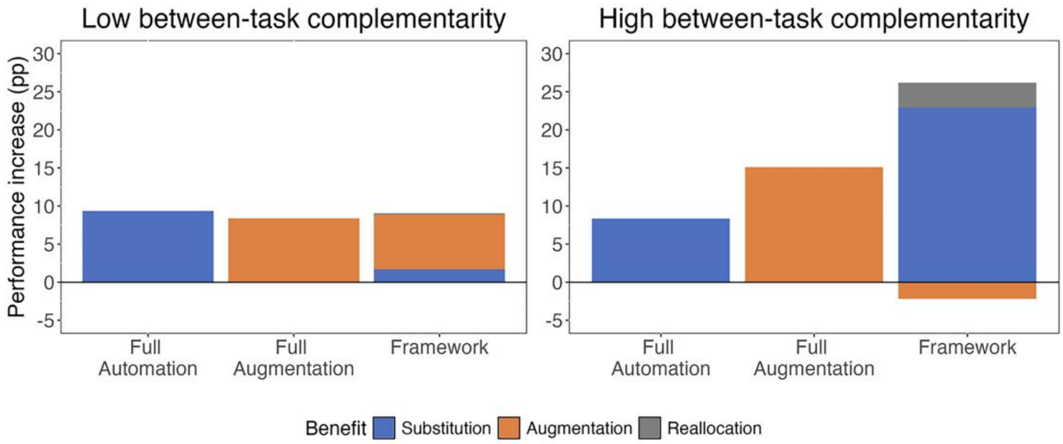
In Figure 6, we show the performance benefits of our framework for images with low within-task complementarity (left panel) and images with high within-task complementarity (right panel). The second decomposition supports Statement 2. Higher within-task complementarity is associated with a lower share of automation benefits (substitution and reallocation) and a higher share of augmentation benefit.

Our third decomposition is designed to explore cases in which AI is, on average, inferior or superior to human performance. The results of this decomposition are shown in Figure 7. In the first group (the left panel of Figure 7), AI performance is inferior to human performance (70% versus 78%). Although full automation is detrimental in this case, our framework achieves an overall performance of 86%. This shows that in a case with both types of complementarity, humans can profit from an inferior AI. The second group (the right panel

Table 1. Summary Statistics for Each Group of Images

Group	$\overline{p^{AI}}$	$\overline{p^H}$	a	b	r^{AI}	$r^{\overline{AI}}$
Low between-task complementarity	0.88	0.79	0.03	0.89	0.76	0.62
High between-task complementarity	0.66	0.58	0.67	−0.13	0.80	0.47
Low within-task complementarity	0.76	0.67	0.44	0.30	0.72	0.61
High within-task complementarity	0.78	0.70	0.45	0.32	0.84	0.41
Inferior AI	0.70	0.78	0.46	0.48	0.77	0.45
Superior AI	0.84	0.58	−0.23	0.94	0.79	0.62
Full data set	0.77	0.68	0.44	0.31	0.78	0.52

Figure 5. (Color online) Benefits of Automation and Augmentation for Low and High Between-Task Complementarity



Note. pp, percentage points.

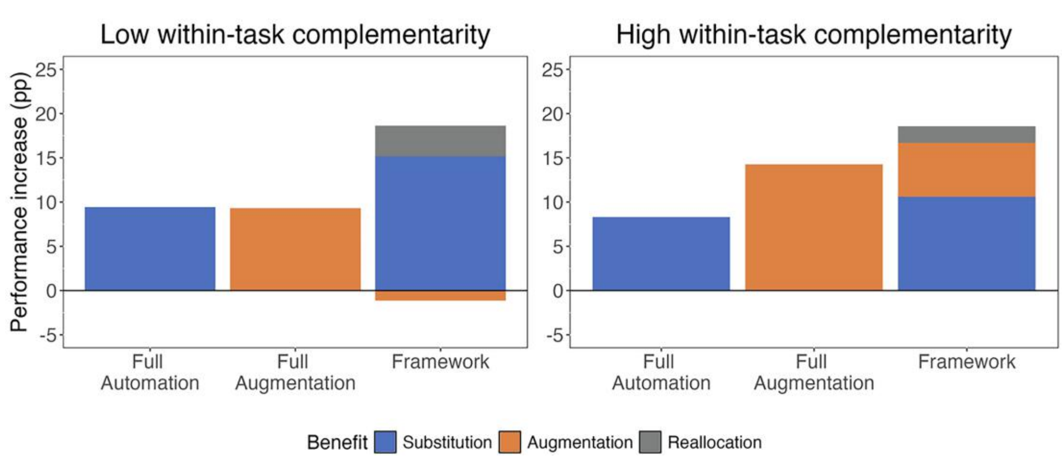
of Figure 7) illustrates a case where AI performance greatly exceeds human performance (84% versus 57%). Here, our framework is not able to surpass the benefit of full automation. This case demonstrates the limited potential of human input if both between-task complementarity and within-task complementarity are of low value. Although this case illustrates a boundary condition for the value of humans within our framework, it also shows the robustness of our framework.

5.3. Distribution of Work

In this section, we analyze the effects on the distribution of work if our task allocation framework is employed. We add the dimension of available human resources to consider possible work distribution patterns when human resources are scarce or plentiful.

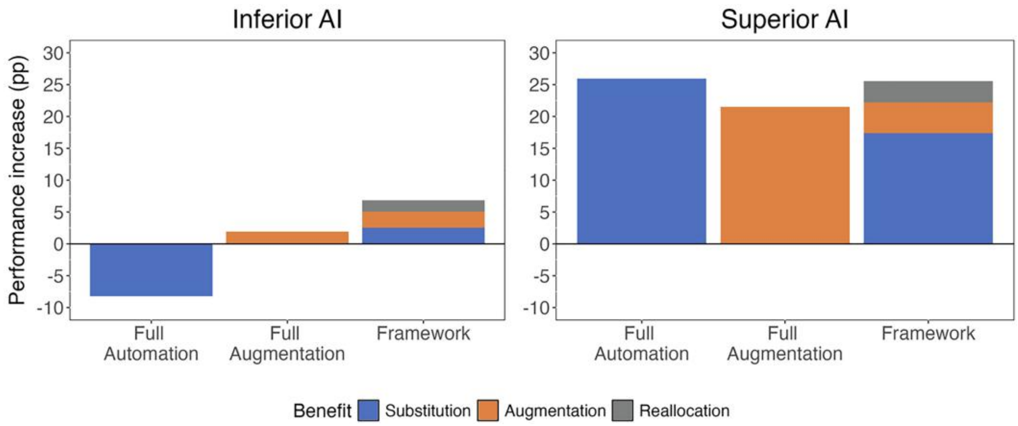
Figure 8 shows the distribution of work for all 100 images sorted in ascending order based on AI certainty on the vertical axis (the first row represents the most difficult image and the last row represents the easiest image from the AI's perspective). The horizontal axis in Figure 8 represents five levels of available human resources, where 100% corresponds to the baseline case where each image is allocated to one human resource. For each human resource level, we illustrate the task allocation results that lead to the maximum expected performance over all images. The results show a consistent pattern; although the images with the highest level of AI certainty are automated by AI, the images with medium AI certainty are performed by humans (individuals or very small groups) augmented with AI, and the images with the lowest level of AI certainty are performed by groups of humans without AI support. For the baseline case with 100% human resources, this leads to 84% of the humans working on the 20% most difficult tasks, increasing their average performance from 59% to 74% in those tasks. Thus, tasks that were formerly perceived to be

Figure 6. (Color online) Benefits of Automation and Augmentation for Low and High Within-Task Complementarity



Note. pp, percentage points.

Figure 7. (Color online) Benefits of Automation and Augmentation for Inferior and Superior AI



Note. pp, percentage points.

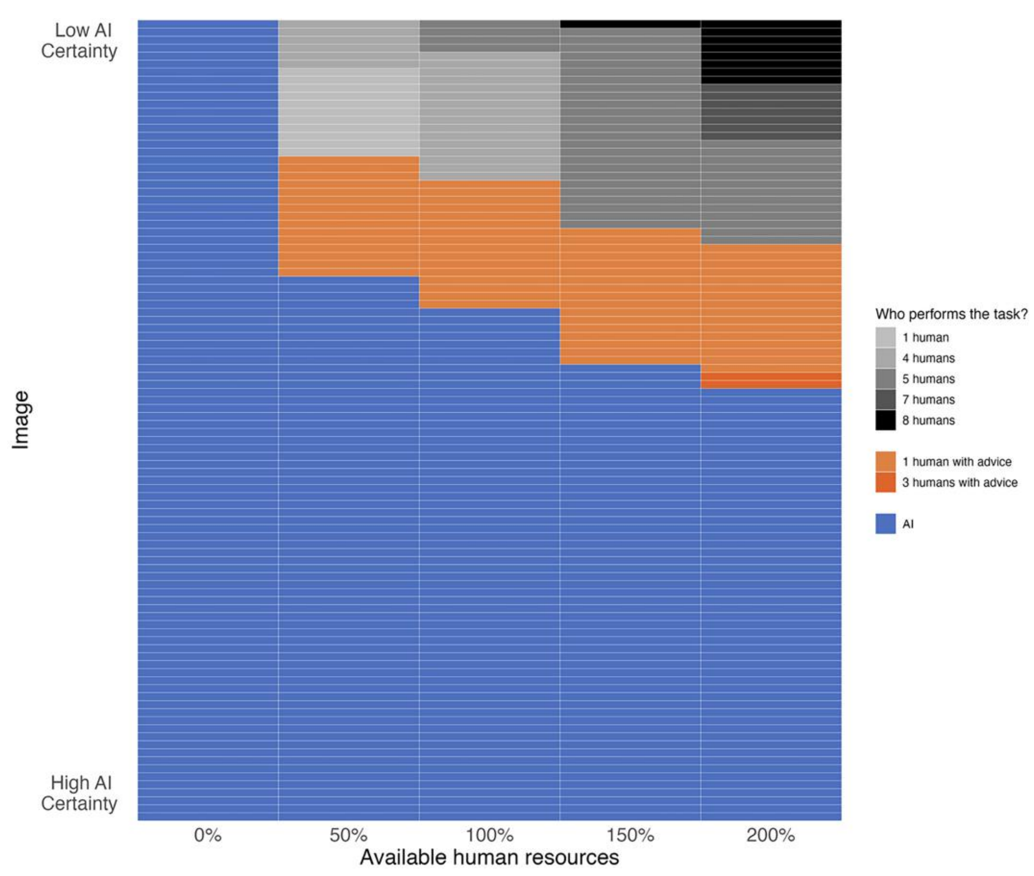
difficult and were solved poorly by humans are now performed with an accuracy that exceeds the average baseline performance of 68%.

We also analyze how the framework deals with additional human resources. Obviously, all work is automated if no human resources are available. Added human resources are allocated to difficult tasks with low AI certainty first, and groups of humans are created. As more human resources become available, AI will increasingly be replaced by groups of human workers.

6. Discussion

The empirical results validate the theoretical statements on the effects of between-task and within-task complementarity on the distribution of work between humans

Figure 8. (Color online) Distribution of Work Between Humans and AI



and AI and on the benefits that AI can provide through automation and augmentation. Based on these results, we derive meaningful theoretical and managerial implications. We believe that these implications hold for any set of (judgment) tasks that satisfy the two conditions specified in Section 3 (the Polanyi condition and the ground truth condition).

6.1. Theoretical Implications

Although the general notion of previous human-AI research seems to be that augmentation is preferable to automation, our analysis suggests that a more nuanced view is necessary to understand the benefits of AI. Our task allocation framework highlights the central tenet of Raisch and Krakowski (2021), who argue that strong interdependencies between the roles of automation or augmentation exist and that these roles, therefore, need to be analyzed simultaneously. Our work demonstrates that the benefits that AI provides through automation and augmentation depend on the type of complementarity between humans and AI.

Our results also corroborate the findings of Dell'Acqua et al. (2023), who argue that the capabilities of AI can be defined along a "jagged technological frontier" (Dell'Acqua et al. 2023, p. 1). For tasks that lie within the frontier (e.g., summarizing ideas), AI can provide benefits by automating the task or by augmenting humans who work on the task. For tasks that lie outside the frontier (e.g., providing strategic recommendations to a company), AI cannot provide any benefits, which means that humans should work on these tasks without AI support.

In our task allocation framework, the introduction of AI can result in different types of benefits compared with the baseline case, where each task is performed by a human without AI. When AI is used for automation, two effects emerge. The *substitution benefit* is a result of AI taking over tasks from humans when AI shows superior performance on these tasks. The *reallocation benefit* results from the fact that the human substituted by AI can be reallocated to another task, which can improve the performance on that task. The share of substitution and reallocation benefits increases for higher levels of between-task complementarity and lower levels of within-task complementarity. When AI is used for augmentation, another benefit arises because the capabilities of the human(s) working on the task are increased (*augmentation benefit*). The share of augmentation benefit increases for higher levels of within-task complementarity and lower levels of between-task complementarity. We argue that a more nuanced differentiation of human-AI complementarity should inform future research on human-AI collaboration.

6.2. Managerial Implications

Our task allocation framework allows us to increase the average accuracy across all tasks with the same amount of (human) resources compared with the accuracy in the baseline case, where humans work alone on the tasks, and compared with the accuracy in the full-automation and full-augmentation benchmarks. Allocating tasks effectively to humans and AI can have strong benefits in terms of quality (i.e., task performance). In our empirical study, the task allocation framework allocates 80% of humans to the 20% most difficult tasks. Because task difficulty is defined from the perspective of AI, difficult tasks might be innovations or unique tasks where little or no data are available. For these tasks with high uncertainty, humans working together can provide performance benefits (Graesser et al. 2018, Samuel et al. 2024). We believe that this could be a scenario in the future of work. AI performs all standardized tasks, where plenty of data are available and AI can be well trained; some humans work with AI advice on more difficult tasks, where AI can still provide some benefits, and most humans will be able to work in groups on novel, difficult, or new tasks, where AI is not (yet) able to provide meaningful input as little to no data exist. Although our framework allocates tasks based on the expected performance of humans and AI as well as their complementarity, we acknowledge that other factors play a crucial role when determining the most appropriate work constellation. For automation, its appropriateness might depend on legal aspects, such as liability. For augmentation, the understanding of humans regarding their own and the AI's capabilities will affect their reliance on the AI's advice. Depending on the nature of the task, other work constellations can be envisioned. For example, if a task can be decomposed into several interdependent subtasks, then humans and AI could divide these subtasks among themselves (Beer et al. 2022).

Applying our framework to the skin cancer detection example from Section 1, this could lead to the following setup. In the first step, all of the cases are evaluated by the AI. All cases with high AI certainty are automated by the AI. In cases with medium AI certainty, physicians review the AI suggestions, and a team of physicians is allocated to cases that are highly uncertain from the AI perspective. In those cases, the physicians who are not required for the automated cases can help improve the detection performance in the most challenging cases.

6.3. Outlook on the Future of AI

In our task allocation framework, the contribution of humans lies in improved performance on tasks where AI may be inferior. The degree to which the human contribution is meaningful depends on the existence of complementarity between humans and AI. Complementarity is a decisive factor that determines whether humans can provide benefits in the future of work. As complementarity is driven by the knowledge gap between humans and AI, it is crucial to critically evaluate whether this knowledge gap will persist in the future of work. We believe that developments in three areas will affect this knowledge gap: (1) improvements in AI, (2) changes in task and process design, and (3) loss of tacit knowledge.

Improvements in AI. The technological frontier of AI capabilities is constantly changing (Dell'Acqua et al. 2023). In the future, the knowledge gap between humans and AI will likely narrow as the AI capabilities increase. Studies in the field of image classification have revealed that although a significant error consistency gap between humans and AI exists (Geirhos et al. 2020), it seems that this gap is becoming narrower as AI becomes more advanced (Geirhos et al. 2021). The extended analysis in Section 5.3 shows a scenario in which AI has very high relative performance. In the "superior AI" group, the performance difference is large as AI achieves an accuracy of 84%, and humans achieve an accuracy of only 58%. The right panel of Figure 7 shows that involving humans in this scenario does not lead to an increase in accuracy. Even though Figure 7 shows that based on ex ante estimates of performance, (augmented) humans are allocated to tasks, the substitution benefit is the main driver of increased accuracy (rather than the augmentation benefit or reallocation benefit). If AI becomes so dominant that humans cannot provide any additional value, we might see the emergence of an algorithmic monoculture, where each task (such as hiring employees) will be performed by only one algorithm across firms (Kleinberg and Raghavan 2021, Bommasani et al. 2022). An alternative view is presented in Acemoglu and Restrepo (2018), who present a dynamic model in which they discuss the future of work under increased levels of automation. They find that although humans are substituted for existing tasks, automation enables the emergence of new tasks where humans have a comparative advantage over AI.

Changes in Task and Process Design. In addition to improvements in AI abilities, environmental control (i.e., the redesign of tasks and processes) is another way to narrow the knowledge gap between humans and AI (Autor 2015). Amaya and Holweg (2024) present two pathways that help to make algorithms more applicable to knowledge work. The *task automation pathway* focuses on changing tasks so that they can be automated by algorithms. The relevant steps include standardizing the task or categorizing task activities in terms of their appropriateness for AI. The *process re-engineering pathway* focuses on changing processes so that algorithms can be employed. The relevant steps include "controlling for process variation, adding steps for assuring algorithmic outputs, and eliminating nonvalue-adding activities" (Amaya and Holweg 2024, p. 15). Whether these pathways will be successful depends on the modularity of the organization that wants to adopt AI (Bresnahan 2021, Agrawal et al. 2024). Both pathways might have the effect that tasks become more formalized and the Polanyi condition would not necessarily be satisfied anymore. In this case, our framework might rely on automation only.

Loss of Tacit Knowledge. If AI improves to the point where it acquires all codified knowledge, humans will only be able to provide complementary benefits through noncodifiable tacit knowledge. An important challenge is, therefore, to determine whether human-AI collaboration affects the tacit knowledge of humans. If tacit knowledge is built only while performing a task (Ryle 1945, Polanyi 1966), then letting humans work with AI will potentially affect their share of tacit knowledge. If the collaboration between humans and AI becomes too close, humans might stop thinking about the task, which could lead to them losing their tacit knowledge (Lebovitz et al. 2021). For example, humans receiving AI advice might experience AI performing reasonably well and thus, may be tempted to blindly follow AI advice without exercising critical judgment. In this way, humans would become mouthpieces of AI and lose their agency (Jussupow et al. 2021). The black-box nature of AI exacerbates this problem because the lack of transparency makes it challenging (if not impossible) for humans to learn from advice and gain further knowledge (Lebovitz et al. 2022, Amaya and Holweg 2024). The deterioration of tacit knowledge would mean that humans would lose a complementary advantage, which could result in increased substitution levels. Therefore, augmentation could lead to more automation over time (see Raisch and Krakowski 2021 for a discussion of these temporal dynamics). Substitution of humans by algorithms would then have detrimental effects on organizational learning outcomes (Balasubramanian et al. 2022). The group with superior AI in Section 5 illustrates a scenario in which the strong performance of AI and the lack of complementarity lead to humans not providing any additional benefits.

6.4. Limitations and Outlook for Future Research

With our task allocation framework, we aim to provide a starting point for discussing the allocation of tasks between humans and AI under different types of complementarity. Many exciting research opportunities exist to extend or adapt the current framework.

The first set of extensions concerns the adaptation of the framework's objective function. In our current version, we maximize the average task performance while keeping human resources (i.e., costs) and the number of tasks fixed. Alternative approaches could choose a different objective; one could minimize the human resources (costs) while keeping performance and the number of tasks fixed. Further, one could maximize the number of completed tasks while keeping performance and human resources fixed. To add more heterogeneity, one could also think about adding "task value" as a performance indicator (in terms of financial outcomes) and optimizing for the product of accuracy and task value. Analyzing the allocation of tasks between humans and AI in terms of task value could reveal interesting allocations, such as that one might sacrifice accuracy on a lower-value task by substituting the human with AI so that human resources could be reallocated to a higher-value task.

The second set of extensions to our study concerns the analysis of tasks with different characteristics. We developed our framework for judgment tasks and highlighted its effectiveness for image classification tasks. An interesting avenue would be to investigate how tasks with different characteristics, such as planning tasks or creativity tasks (McGrath 1984), change the dynamics between humans and AI. This aspect is crucial because different kinds of AI algorithms, such as reinforcement learning algorithms (Sutton and Barto 2018) and large language models (Dell'Acqua et al. 2023), are employed in these tasks. An example of how our logic might transfer to future work configurations in creative tasks is journalism; AI could write simple announcements, human journalists could work with AI on regular reports, and groups of journalists could work on long-term investigative assignments.

The third set of extensions concerns the modeling of the human workers. Here, one could add heterogeneity of humans by considering observable differences in task performance as well as differences in their abilities to differentiate correct and incorrect advice. Although the former would allow us to model the impact of human heterogeneity on between-task complementarity, the latter would allow us to model its impact on within-task complementarity. Another extension could be to relax the assumption of independence between crowd members because in reality, the errors of the crowd members might be correlated (Palley and Soll 2019). Further, although we treat humans as freely (re-)assignable between tasks in our framework, another extension could be to consider task switching costs. Because AI substitutes humans for easy tasks, humans are left with the medium and difficult tasks. This reallocation could involve some inefficiencies as humans would need to get accustomed to these new tasks.

6.5. Conclusion

It is very exciting to witness the ever-growing importance of AI in the discussion of the future of work. With this paper, we contribute to this discussion by theoretically outlining the benefits that AI can provide when collaborating with humans. Our task allocation framework and our empirical analyses show that the benefits of AI depend on the existence of between-task and within-task complementarity. AI can support humans in many capacities by taking over tasks, augmenting humans, or allocating humans to tasks where AI cannot provide benefits. In summary, humans and AI will collaborate closely, and the division of labor and their respective contributions will be heavily influenced by how well they complement each other.

References

Acemoglu D, Restrepo P (2018) The race between man and machine: Implications of technology for growth, factor shares, and employment. *Amer. Econom. Rev.* 108(6):1488–1542.

Agrawal A, Gans J, Goldfarb A (2018) *Prediction Machines: The Simple Economics of Artificial Intelligence* (Harvard Business Press, Cambridge, MA).

Agrawal A, Gans J, Goldfarb A (2024) AI adoption and system-wide change. *J. Econom. Management Strategy* 33(2):327–337.

Amaya J, Holweg M (2024) Using algorithms to improve knowledge work. *J. Oper. Management* 70(3):482–513.

Asch SE (1955) Opinions and social pressure. *Sci. Amer.* 193(5): 31–35.

Autor DH (2014) *Polanyi's Paradox and the Shape of Employment Growth* (National Bureau of Economic Research, Cambridge, MA).

Autor DH (2015) Why are there still so many jobs? The history and future of workplace automation. *J. Econom. Perspect.* 29(3):3–30.

Autor DH, Levy F, Murnane RJ (2003) The skill content of recent technological change: An empirical exploration. *Quart. J. Econom.* 118(4):1279–1333.

Baird A, Maruping LM (2021) The next generation of research on IS use: A theoretical framework of delegation to and from agentic IS artifacts. *MIS Quart.* 45(1):315–341.

Balasubramanian N, Ye Y, Xu M (2022) Substituting human decision-making with machine learning: Implications for organizational learning. *Acad. Management Rev.* 47(3):448–465.

Bansal G, Nushi B, Kamar E, Lasecki WS, Weld DS, Horvitz E (2019) Beyond accuracy: The role of mental models in human-AI team performance. *Proc. AAAI Conf. Human Comput. Crowdsourcing* 7(1):2–11.

Beam AL, Kohane IS (2016) Translating artificial intelligence into clinical care. *JAMA* 316(22):2368–2369.

Beer R, Qi A, Ríos I (2025) Behavioral externalities of process automation. *Management Sci.*, ePub ahead of print July 2, https://doi.org/10.1287/mnsc.2024.05408.

Blattberg RC, Hoch SJ (1990) Database models and managerial intuition: 50% model + 50% manager. *Management Sci.* 36(8): 887–899.

Boland PJ, Proschan F, Tong YL (1989) Modelling dependence in simple and indirect majority systems. *J. Appl. Probab.* 26(1): 81–88.

Bommasani R, Creel KA, Kumar A, Jurafsky D, Liang PS (2022) Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? *Adv. Neural Inform. Processing Systems* 35:3663–3678.

Bonaccio S, Dalal RS (2006) Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organ. Behav. Human Decision Processes* 101(2):127–151.

Bresnahan TF (2021) Artificial intelligence technologies and aggregate growth prospects. Diamond JW, Zodrow GR, eds. *Prospects for Economic Growth in the United States* (Cambridge University Press, Cambridge, UK), 132–172.

Broussard M (2018) *Artificial Unintelligence: How Computers Misunderstand the World* (MIT Press, Cambridge, MA).

Brynjolfsson E, Mitchell T (2017) What can machine learning do? Workforce implications. *Science* 358(6370):1530–1534.

Brynjolfsson E, Mitchell T, Rock D (2018) What can machines learn and what does it mean for occupations and the economy? *AEA Papers Proc.* 108:43–47.

Chekroud AM, Hawrilenko M, Loho H, Bondar J, Gueorguieva R, Hasan A, Kambeitz J, et al. (2024) Illusory generalizability of clinical prediction models. *Science* 383(6679):164–167.

Chen C, Jain N, Karamshetty V (2023) Algorithm-human-algorithm: A new classification approach to integrating judgemental adjustments. *Proc. 44th Internat. Conf. Inform. Systems (Hyderabad, India)*.

Condorcet M D (1785) *Essay on the Application of Analysis to the Probability of Majority Decisions* (Imprimerie Royale, Paris).

Dell'Acqua F, McFowland E, Mollick ER, Lifshitz-Assaf H, Kellogg K, Rajendran S, Krayer L, Candelon F, Lakhani KR (2023) Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. Working paper, Harvard Business School Technology & Operations Management Unit, Cambridge, MA.

Dellermann D, Ebel P, Söllner M, Leimeister JM (2019) Hybrid intelligence. *Bus. Inform. Systems Engrg.* 61(5):637–643.

Donahue K, Chouldechova A, Kenthapadi K (2022) Human-algorithm collaboration: Achieving complementarity and avoiding unfairness. *Proc. 2022 ACM Conf. Fairness Accountability Transparency (Seoul, South Korea)*, 1639–1656.

Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S (2017) Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542(7639):115–118.

Frank MR, Autor D, Bessen JE, Brynjolfsson E, Cebrian M, Deming DJ, Feldman M, Groh M, Lobo J, Moro E (2019) Toward understanding the impact of artificial intelligence on labor. *Proc. Natl. Acad. Sci. USA* 116(14):6531–6539.

Fügener A, Grahl J, Gupta A, Ketter W (2021) Will humans-in-the-loop become borgs? Merits and pitfalls of working with AI. *MIS Quart.* 45(3):1527–1556.

Fügener A, Grahl J, Gupta A, Ketter W (2022) Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation. *Inform. Systems Res.* 33(2): 678–696.

Geirhos R, Meding K, Wichmann FA (2020) Beyond accuracy: Quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. *Adv. Neural Inform. Processing Systems* 33:13890–13902.

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

Graesser AC, Fiore SM, Greiff S, Andrews-Todd J, Foltz PW, Hesse FW (2018) Advancing the science of collaborative problem solving. *Psych. Sci. Public Interest* 19(2):59–92.

Hekler A, Utikal JS, Enk AH, Hauschild A, Weichenthal M, Maron RC, Berking C, et al. (2019) Superior skin cancer classification by the combination of human and artificial intelligence. *Eur. J. Cancer* 120:114–121.

Jain H, Padmanabhan B, Pavlou PA, Raghu TS (2021) Editorial for the special section on humans, algorithms, and augmented intelligence: The future of work, organizations, and society. *Inform. Systems Res.* 32(3):675–687.

Jussupow E, Spohrer K, Heinzl A, Gawlitza J (2021) Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence. *Inform. Systems Res.* 32(3):713–735.

Karush W (1962) A generalalgorithm for the optimal distribution of effort. *Management Sci*. 9(1):50–72.

Kesavan S, Kushwaha T (2020) Field experiment on the profit implications of merchants' discretionary power to override data-driven decision-making tools. *Management Sci.* 66(11): 5182–5190.

Kleinberg J, Raghavan M (2021) Algorithmic monoculture and social welfare. *Proc. Natl. Acad. Sci. USA* 118(22):e2018340118.

Lake BM, Salakhutdinov R, Tenenbaum JB (2015) Human-level concept learning through probabilistic program induction. *Science* 350(6266):1332–1338.

Lake BM, Ullman TD, Tenenbaum JB, Gershman SJ (2017) Building machines that learn and think like people. *Behav. Brain Sci.* 40:e253.

Lebovitz S, Levina N, Lifshitz-Assaf H (2021) Is AI ground truth really true? The dangers of training and evaluating AI tools based on experts' know-what. *MIS Quart.* 45(3):1501–1525.

Lebovitz S, Lifshitz-Assaf H, Levina N (2022) To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Organ. Sci.* 33(1):126–148.

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521(7553):436–444.

Lindebaum D, Vesa M, Den Hond F (2020) Insights from "the machine stops" to better understand rational assumptions in algorithmic decision making and its implications for organizations. *Acad. Management Rev.* 45(1):247–263.

List C, Goodin RE (2001) Epistemic democracy: Generalizing the Condorcet jury theorem. *J. Political Philos.* 9(3):277–306.

Madras D, Pitassi T, Zemel R (2018) Predict responsibly: Improving fairness and accuracy by learning to defer. *Adv. Neural Inform. Processing Systems* 31:6150–6160.

Maedche A, Legner C, Benlian A, Berger B, Gimpel H, Hess T, Hinz O, Morana S, Söllner M (2019) AI-based digital assistants. *Bus. Inform. Systems Engrg.* 61(4):535–544.

McGrath JE (1984) *Groups: Interaction and Performance* (Prentice-Hall, Englewood Cliffs, NJ).

Michelucci P, Dickinson JL (2016) The power of crowds. *Science* 351(6268):32–33.

Minsky M, Papert S (1969) *Perceptrons* (MIT Press, Cambridge, MA).

Mozannar H, Sontag D (2020) Consistent estimators for learning to defer to an expert. *Proc. 37th Internat. Conf. Machine Learn.* (JMLR.org), 7076–7087.

Orlikowski WJ (2002) Knowing in practice: Enacting a collective capability in distributed organizing. *Organ. Sci.* 13(3):249–273.

Palley AB, Soll JB (2019) Extracting the wisdom of crowds when information is shared. *Management Sci.* 65(5):2291–2309.

Polanyi M (1966) The logic of tacit inference. *Philosophy* 41(155):1–18.

Puranam P (2021) Human–AI collaborative decision-making as an organization design problem. *J. Organ. Design* 10(2):75–80.

Raghu M, Blumer K, Corrado G, Kleinberg J, Obermeyer Z, Mullainathan S (2019) The algorithmic automation problem: Prediction, triage, and human effort. Preprint, submitted March 28, https://arxiv.org/abs/1903.12220.

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

Raisch S, Krakowski S (2021) Artificial intelligence and management: The automation–augmentation paradox. *Acad. Management Rev.* 46(1):192–210.

Rosenfeld A, Solbach MD, Tsotsos JK (2018) Totally looks like-how humans compare, compared to machines. *Proc. IEEE Conf. Comput. Vision Pattern Recognition Workshop* (IEEE Computer Society, Washington, DC), 1961–1964.

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M (2015) ImageNet large scale visual recognition challenge. *Internat. J. Comput. Vis.* 115(3):211–252.

Ryle G (1945) Knowing how and knowing that: The presidential address. *Proc. Aristotelian Soc.* 46(1):1–16.

Samuel J, Zheng Z (E), Mookerjee V (2024) Task characteristics and incentives in collaborative problem solving: Evidence from three field experiments. *Inform. Systems Res.* 35(1): 414–433.

Schuetz S, Venkatesh V (2020) The rise of human machines: How cognitive computing systems challenge assumptions of user-system interaction. *J. Assoc. Inform. Systems* 21(2):460–482.

Simon HA (1957) *Models of Man; Social and Rational* (Wiley, New York).

Smith BC (2019) *The Promise of Artificial Intelligence: Reckoning and Judgment* (MIT Press, Cambridge, MA).

Steyvers M, Tejeda H, Kerrigan G, Smyth P (2022) Bayesian modeling of human–AI complementarity. *Proc. Natl. Acad. Sci. USA* 119(11):e2111547119.

Sutton RS, Barto AG (2018) *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA).

Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. *Proc. IEEE Conf. Comput. Vision Pattern Recognition (Las Vegas, NV)*, 2818–2826.

Tschandl P, Rosendahl C, Akay BN, Argenziano G, Blum A, Braun RP, Cabo H, et al. (2019) Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks. *JAMA Dermatology* 155(1):58–65.

Tversky A, Kahneman D (1974) Judgment under uncertainty: Heuristics and biases. *Science* 185(4157):1124–1131.

Walzner DD, Poreschack LM, Fuegener A, Schiffels S, Denz C (2023) Division of labor between humans and algorithms in healthcare: The case of surgery duration predictions. *Proc. 44th Internat. Conf. Inform. Systems (Hyderabad, India).*

Yaniv I, Kleinberger E (2000) Advice taking in decision making: Egocentric discounting and reputation formation. *Organ. Behav. Human Decision Processes* 83(2):260–281.