December 2024

Ensuring Human Agency: A Design Pathway to Human-AI Interaction

Anuschka Schmitt
The London School of Economics and Political Science, a.schmitt2@lse.ac.uk

Ensuring Human Agency: A Design Pathway to Human-AI Interaction

Completed Research Paper

Anuschka Schmitt
The London School of Economics and Political Science
Houghton Street, London WC2A 2AE
a.schmitt2@lse.ac.uk

Abstract

The augmentation of human work through Artificial Intelligence (AI) promises to be a panacea to the role of technology in organizations. While frameworks on augmentation theorize how to best divide work between humans and AI, the empirical literature on human-AI interaction offers unexpected and inconclusive findings. Interaction challenges—including overreliance and selected engagement with the algorithmic output—call into question how theorized augmentation benefits can be realized. Rooted in cognitive learning theory, this study's conceptual model argues that human-AI interaction can lead to multiple beneficial outcomes when algorithmic output is designed in a reciprocal manner. By providing humans with reflection-provoking feedback, reciprocal algorithmic output does not prescribe any actions, and thereby necessitates humans to expend cognitive effort. Reciprocal algorithmic output enables three crucial augmentation outcomes: task performance, human agency, and human learning.

Keywords: human-AI interaction, augmentation, algorithmic output, cognitive learning, agency

Introduction

It is hard to target human-complementary work in the abstract.

- *Acemoglu, Autor, and Johnson, 2023, 'Can We Have Pro-Worker AI?', p.8*

By leveraging the capabilities of the human and the AI, augmentation promises to enable productivity gains neither AI nor human would be capable of achieving on their own (Rai et al., 2019). The potential of human and AI working together appears to be more important than ever: while AI has been esteemed for its superior prediction accuracy and lack of human bias, its generation of incorrect and biased output remains largely inscrutable (Berente et al., 2021). The opaque and unpredictable nature of AI thereby not only distinguishes AI from previous technology yet requires a thorough understanding of how to leverage AI for organizational benefit. Research on AI in information systems (IS) and management has theorized how to divide work between AI and humans (Baird & Maruping, 2021; Teodorescu et al., 2021), and how such division of work would look like for certain tasks (Fügener et al., 2021; Murray et al., 2021).

Empirical findings suggest that the hoped-for performance benefits of augmentation are difficult to realize, though. Algorithmic output does not necessarily improve, yet often even worsens, performance outcomes. Humans struggle to assess when to rely on algorithmic output and thereby rely on incorrect output, too (Jussupow et al., 2021; Logg et al., 2019). Humans also seem to engage with algorithmic output only selectively, i.e., when it confirms their preconceptions (Abdel-Karim et al., 2023; Bauer et al., 2023; Lebovitz et al., 2022).

Insufficient attention is being given to the design of human-AI interaction. While empirical studies on AI augmentation have explored important cognitive mechanisms of why humans rely on such output, they follow a prevalent design of algorithmic output: an outcome-focused recommendation of one (best) solution. Conceptual frameworks suggest crucial ontological ideas of how tasks can be shared between AI and humans yet treat the design of algorithmic output as given. Paying explicit attention to the design of human-AI interaction offers a promising path for overcoming previously mentioned interaction challenges, and realizing beneficial augmentation outcomes.

Furthermore, the embedding of human-AI interaction within organizational contexts demands to consider the organizational boundary conditions as well as the diverse and long-term implications of augmentation (Parker & Grote, 2022; Zuboff, 1985). To date, much of the discussion has emphasized potential performance gains. It is unclear how AI and its output can enable and balance a multiplicity of beneficial augmentation outcomes. This becomes particularly important in organizational contexts where we might aim to improve performance without forgoing human agency and skill development.

The design of algorithmic output might be ideally positioned to reorient the current debates of AI augmentation towards a more proactive stance on what forms of human-AI interactions are desirable and how we can get there. Relying on phenomenon-driven theorizing (Fisher et al., 2021), this study draws on learning theory to develop a conceptual model on human-AI interaction complementing and extending prior work on AI augmentation. Our conceptual model suggests that the design of algorithmic output holds a central role in understanding the effects of human-AI interaction. Accordingly, human-AI interactions can lead to a beneficial augmentation of human work if algorithmic output is framed in a reciprocal manner. Reciprocal algorithmic output provides open-ended, reflection-provoking feedback. Output thereby integrates a user's input, and the user must expend effort to arrive at an answer. *Reciprocal algorithmic output*

- provides the user with evaluative feedback and critique (as compared to an explicit, outcome-focused recommendation) and
- focuses on improving a user's understanding of the augmented task and domain (as compared to improving users' understanding of the AI-based system).

Reciprocal algorithmic output thus serves as a supportive, rather than a prescriptive, function for AI augmentation. Our model for human-AI interaction conceptualizes three crucial augmentation outcomes that reciprocal algorithmic output enables:

- *Task performance* is concerned with productivity gains, a predominant interaction outcome paid attention to in the literature. With the commercial availability and processing capabilities of large language model-based systems, AI-based systems offer increasing efficacy, e.g., in terms of accuracy, quality, or speed of task completion (e.g., Noy & Zhang, 2023).
- *Human agency* refers to the degree of autonomy and control humans exercise and has positive implications for workers' motivation and self-determination (Hackman & Oldham, 1976). In an augmentation context, human agency might be (unintentionally) limited to having humans reject or accept algorithmic output (Murray et al., 2021) or having humans delegate (sub)tasks to AI (Fügener et al., 2021).
- *Human learning* is concerned with humans maintaining and enhancing their skills. Learning enables humans to remain capable of executing tasks without AI, as well as assessing AI-based systems' output. A longitudinal study by Abbas et al. (2024), for instance, illustrated how university students' usage of ChatGPT was linked to procrastination and difficulties in remembering information, as well as decreasing performance over time.

Reciprocal algorithmic output leads to increased task performance by overcoming current interaction challenges. By providing pro and contra arguments for a user's judgement, for instance, reciprocal output is not making a conclusive recommendation and, in turn, cannot mislead the user to follow incorrect output. Conversely, reciprocal algorithmic output ensures human agency as an outcome must be developed by the human. By design, the user is not able to forgo their agency by simply accepting the output. Finally, reciprocal algorithmic output leads to more human learning, since the output necessitates the user to expend cognitive effort.

A design-focused, sociotechnical perspective on human sensemaking of algorithmic output complements macro-level theorizing and ontological debates of how to balance AI and human agency in augmentation scenarios. Our conceptualization thereby follows IS design theories motivated by real-world phenomena and rooted in selected kernel theories (e.g., Kane et al., 2021). Our work acknowledges key interaction challenges in AI augmentation while theorizing how organizations can overcome these challenges to better realize beneficial forms of human-AI interaction.

Conceptual Background

Before presenting the conceptual model, we summarize the relevant prior work on the theoretical notions of and empirical findings on augmentation.

Theoretical Perspectives on AI Augmentation

With the commercialization and scalability of transformer- and large language model-based systems, non-routine and knowledge intensive jobs are no longer immune to the threat of automation (Acemoglu et al., 2023). Accordingly, multiple theoretical frameworks have been introduced in the past few years, recognizing systems' increasing capabilities (Baird & Maruping, 2021; Murray et al., 2021). As such, systems can vary in their agency from reacting to pre-defined stimuli towards being fully autonomous.

While automation can make sense in certain cases, e.g., to replace monotonous or dangerous work (Walsh & Strano, 2018), several considerations call into question the validity of headline predictions around workforce automation (e.g., Frey & Osborne, 2017). For one, in many cases, automation has been shown to only lead to marginal increases in productivity while exacerbating inequality among workers (Acemoglu & Restrepo, 2022). Two, as pointed out by Parker and Grote (2022, p.1172), "tasks exist within a broader role alongside other tasks". It appears unlikely that whole jobs versus individual tasks will be automated. Third, and maybe most importantly, an excessive focus on automation neglects important considerations of technology's usefulness for human work in the long-term, e.g., workers' skill development and sense of purpose (Acemoglu and Johnson, 2023; Zuboff, 1985).

Augmentation aims to create synergies between humans and AI as they jointly work together. Such a joint optimization approach promises to leverage benefits and advantages both human and AI offer while reducing limitations and risks human and AI each also bring along (Rai et al., 2019; Zagalsky et al., 2021). Murray et al.'s (2021) framework on conjoined agency defines an augmenting technology to provide the human with specific recommendations. It is then up to the human to decide whether to follow these recommendations. Focusing on non-routine, exploratory tasks, Raisch and Fomina (2024, p.14) follow a similar conceptualization of AI augmentation. The authors suggest that humans and AI engage in "interactive selection" by having the human select a final decision in consideration of the AI's recommendation. This interaction, however, assumes that humans are "likely to form independent opinions of and preferences for jointly developed solutions, but [are] unlikely to disregard AI's anticipatory quantification completely" (Raisch & Fomina, 2024, p.14).

Much has been said regarding the division of decision-making locus between human and AI in augmentation scenarios. However, little attention has been given to how such augmentation is operationalized or designed, especially in consideration of the sociotechnical nature of human-AI interactions. Questioning these frameworks underlying—explicit and implicit—assumptions might be relevant in order to understand and address why theorized augmentation benefits do not seem to materialize. Murray et al. (2021, p.558) point towards interaction challenges that can arise when previously mentioned assumptions do not hold: "If humans do not consider relevant circumstances when selecting an action, but blindly follow an augmenting technology's recommendation, suboptimal or inappropriate action selection is possible."

At the same time, several theory papers point towards the challenge of clearly defining AI augmentation and its desirable outcomes. Teodorescu et al.'s (2021) typology, for instance, focuses on achieving fairness in augmentation scenarios. Rai et al. (2019) discuss different models of human-AI interaction and introduce the example of industrial workers wearing robotic devices which enable them to exercise their work with more strength. As such, AI and humans are not viewed as equal parts. Rather, AI is viewed to complement the human. This not only provides a crucial distinction of different augmentation conceptualizations, that

is, dividing tasks between AI and human versus AI taking a supportive function. More so, the example of industrial workers points towards relevant augmentation goals beyond productivity gains: while the robotic devices augment workers in their strength, the workers remain control over the tasks.

This notion of human-AI interaction is congruent with seminal conceptualizations of using digital artifacts to augment humans' cognitive abilities: Douglas Engelbart's (1962) conceptual framework on 'Augmenting human intellect' and JCR Licklider's (1960) seminal paper 'Man-Computer-Symbiosis' were not focused on short-term performance gains yet in how to strengthen workers' skills and work quality by augmenting human capabilities. In a similar vein, Zuboff (1985) suggested two strategies with respect to how the output of 'intelligent technology' could be designed. When output is constructed in a way that replaces human effort and skill, an organization rather follows an 'automate' strategy. Her idea of an 'informate' strategy implied to use "the information generated by the automated processes to provide feedback to workers, who are then empowered to make complex decisions" (Parker & Grote, 2022, p.1185).

As systems are becoming increasingly autonomous and applicable to different types of tasks, theoretical frameworks have brought forward novel ideas of how tasks can be worked on jointly by humans and AI. Literature provides distinct views of augmentation, including delegation (Baird & Maruping, 2021), focus on particular types of tasks (Murray et al., 2021; Raisch & Fomina, 2024), and augmentation outcomes beyond productivity gains (Teodorescu et al., 2021). While these frameworks identify the capabilities and mechanisms of humans and AI each, literature has yet to explicitly address how augmentation can be realized from a sociotechnical perspective, i.e., is to be designed in human-AI interaction contexts. Building on and extending seminal notions of augmentation (Engelbart, 1962; Licklider, 1960; Zuboff, 1985), we also need to recognize how such increasingly agentic systems modify the role of humans in our interactions with these systems.

Empirical Findings on Human-AI Interaction

Beyond theoretical notions of AI augmentation, empirical studies have explored human-AI interaction scenarios. Two dominant findings point towards the idea that the way humans and AI interact do not lead to the desired outcomes theoretical frameworks on AI augmentation propose (see Table 1 for an overview).

For one, humans have difficulties to assess when and how to appropriately rely on AI (Fügener et al., 2022; Lebovitz et al., 2022). Experimental and qualitative interview studies find converging evidence that humans overrely on algorithmic output (e.g., Lebovitz et al., 2023; Logg et al., 2019). While accurate output seems to only marginally increase performance, erroneous output worsens performance significantly (Jussupow et al., 2021). This phenomenon of overreliance, i.e., humans also following incorrect output, illustrates that reliance on algorithmic output is not always appropriate, and can worsen performance. Incorrect or fabricated algorithmic output is not predictable yet common, as the stochastic nature of machine learning models makes them sensitive to slight deviations in data (Townsend et al., 2024).

Empirical findings point towards a second common behavioral pattern in human-AI interactions: when output would deviate from human judgement, humans were not only prone to make erroneous decisions yet also engaged less with the output (Bauer et al., 2023; Jussupow et al., 2021). In another medical study context, humans reflected more critically or shallowly on a decision depending on whether the system's output was congruent with their judgement (Abdel-Karim et al., 2023). Algorithmic output appears to be used in unintended ways, i.e., to confirm humans' preconceptions rather than reflecting on one's judgement.

Interaction challenge	Design of algorithmic output	Related computational challenges	Implications for human agency	Example
Overreliance: Easily induced to follow output (even if it is incorrect)	Recommendation of unequivocal, 'one best' solution	Unpredictability: algorithmic output can be incorrect and fabricated		Anchoring bias: humans unquestioningly follow the recommendation

Selected engagement: Output is only considered when it is congruent with human judgement	Framed the same way as task outcome / outcome of human's own sensemaking		Limited autonomy as humans a) can limit effort to confirming / rejecting the output b) become fixated on the recommendation	Confirmation bias: humans only engage with the output when it confirms their (potentially incorrect) judgement, so that no new or contrasting input is considered
Explanations exacerbate previous two issues	Provides information on underlying model and evidence for how model arrived at its recommendation	Opacity: model's prediction or generation process can be inscrutable	Limited control as humans do not necessarily improve their understanding of the task or domain	Signaling effect: humans view explanation as a cue for trustworthiness and reliability

Table 1. Interaction Challenges in AI Augmentation

Explanations seem to exacerbate the two previously identified interaction challenges of overreliance and selected engagement (Gajos & Mamykina, 2022; Rudin, 2019). Explanations serve to address issues of opacity associated with contemporary AI-based systems by making more transparent the underlying models of these systems and how they arrive at their output (Arya et al., 2019). For instance, explanations provide humans with knowledge about the features of a model. However, empirical findings suggest that explanations do not enable humans to rely on correct and refrain from incorrect algorithmic output (Schoeffer et al., 2024) and can decrease task performance (Bauer et al., 2023). In line with other empirical findings, the use of explanations led to an unconsidered confirmation bias rather than helping humans improve their decision quality. It is unclear whether system-focused explanations enable humans, in particular laypeople, to make informed decision.

Taken together, the unexpected interaction challenges found in AI augmentation studies appear to be closely linked to underlying assumptions of how humans make sense of algorithmic output. It is thereby worthwhile to consider how algorithmic output is commonly designed. AI is commonly conceptualized as an 'expert' in that it provides the user with accurate and helpful advice, which in turn, allows the user to improve their performance. Algorithmic output is thereby operationalized as outcome-focused feedback, i.e., by recommending the correct answer (e.g., 'disease' or 'no disease' in a medical decision-making context, Jussupow et al., 2021), or indicating accuracies for multiple predictions (e.g., Abdel-Karim et al., 2023; Zagalsky et al., 2021).

The effectiveness of such algorithmic output and reviewed theoretical frameworks rest on two crucial underlying assumptions that do not seem to hold: one, that the output of the AI is predictable, i.e., in that it provides the human continuously with useful and accurate information. Two, that the human is in control, i.e., in that the human can assess when to rely on correct output and when to refrain from incorrect output (Table 1 illustrates how these assumptions are challenged and do not seem to hold in empirical findings).

More so, it appears that empirical studies so far have focused on efficacy-related outcomes such as accuracy or speed of decision-making. While performance and quality are crucial to decision-making and other organizational tasks, previously reviewed theoretical notions on AI augmentation point towards complementary outcomes relevant to the augmentation of human work (Engelbart, 1962; Licklider, 1960; Teodorescu et al., 2021).

An Interaction Design Perspective for AI Augmentation

For automation scenarios, AI itself automates human work. For augmentation scenarios, AI itself does not augment human work. AI augmentation necessitates a human-AI interaction, and therefore hinges on

human perception and sensemaking of the algorithmic output. Beyond ensuring that AI-based systems are accurate, reliable, and trustworthy, it is how AI is used and how the interaction with AI is designed that matters for augmentation to succeed. That is, we need to pay attention to how to design the output of and the interaction with an AI-based system when being deployed to augment human work.

This study places the design of algorithmic output at the heart of understanding and shaping human-AI interaction. Rooted in a sociotechnical system perspective, there is a sound body of knowledge considering work practices and human sensemaking when designing systems, and related interactions (e.g., Gregor & Benbasat, 1999; Te'eni, 2001). Our study's methodological approach joins IS design theories (e.g., Dhaliwal & Benbasat, 1996; Kane et al., 2021). This study's focus on the design of human-AI interaction is phenomenon-driven and motivated by the unintended and unconsidered consequences of algorithmic output in augmentation scenarios. Learning literature (Jörg, 2004; Vygotsky, 1978; King, 1990; King, 1995) and work on human reasoning and cognition (Alvesson & Spicer, 2012; Carroll & McKendree, 1987) form the conceptual foundations of our design theory.

The Role of Output Design for Human-AI Interaction

We develop a theory of reciprocal output to explain how organizations and system designers can overcome interaction challenges in AI augmentation. Our review on empirical findings of human-AI interaction shows that 1) humans can be easily persuaded of algorithmic output and that 2) persuasion depends on whether algorithmic output aligns with human judgement. Potentially, we can overcome these challenges by designing output in a way that does not try to persuade the human, e.g., by rather encouraging them to reflect. We can also overcome these challenges by designing output in a way so it does not become a matter of (in)congruence with human judgement, e.g., by not providing an explicit recommendation. Figure 1 shows our conceptual model for explaining the role of reciprocal algorithmic output for beneficial augmentation outcomes in human-AI interaction. Human-AI interaction is manifested in the direct relationship between the design of algorithmic output and the augmentation outcomes by the human expending effort on making sense of the output.

The following boundary assumptions help understand the model: first, the increasingly conversational and computational nature of AI-based systems make these systems capable of positively augmenting human work, at least in theory (Murray et al., 2021; Townsend et al., 2023). This is because AI-based systems can act with increasing autonomy, situatedness, and flexibility (Baird & Maruping, 2021; Gregory et al., 2021). Applying these capabilities to an augmentation scenario, AI-based systems should be able to provide situated, real-time output that is personalizable to the user. As such, AI-based systems are aware of user behavior and capable of referring to and integrating user input. Ultimately, these system capabilities imply that *organizations see a strategic benefit or opportunity in AI augmentation*, e.g., as compared to automation (Assumption 1).

Second, this study understands AI augmentation to have individual level implications beyond the strategic, organizational benefits (e.g., Gregory et al., 2021; Kemp, 2023). If task performance were to be the only and most important outcome of AI augmentation scenarios, it becomes questionable why task execution is not dealt with as a tradeoff of human versus AI accuracy (as we would see with automation and delegation scenarios). As put forward by Parker and Grote (2022, p.1174): "Even with more agentic and automated technical systems […] much work will entail an intense interaction between humans and self-learning autonomous technology." An augmentation scenario therefore necessitates that *an AI-based system directly converses with the human* (Assumption 2).

Third, human sensemaking in augmentation scenarios appears to be congruent with general human reasoning behavior (in organizational contexts). If humans interact with AI-based systems in the context of organizational tasks and work processes, actions and decisions are probably bound by time pressure and organizational expectations of using deployed AI-based systems. Functional stupidity, humans' "unwillingness and a (learned) incapacity to engage in reflexivity [and] intellectual effort" (Alvesson & Spicer, 2012, p.1213), becomes a relevant lens for interpreting the interaction challenges we see in AI augmentation scenarios: humans were not seeking to reflect or justify their decision in a substantially engaging manner but focused on the efficient completion of a task, especially if algorithmic output was congruent with their initial judgement. Accordingly, we assume that *users will not expend effort* unless the interaction is constructed in a way that forces users to do so (Assumption 3).

The Multiplicity of Augmentation Outcomes

A core part of our theorizing involves accounting for the multiplicity of important goals of augmentation. We build on prior research that has viewed human agency and skill development as important augmentation goals beyond productivity gains (Engelbart, 1962; Licklider, 1960; Zuboff, 1985). Augmentation benefits both the organization which is deploying AI as well as the worker whose work is being augmented (see Assumptions 1 and 2). In this section, we elucidate three key outcomes that are crucial to the core idea of augmentation: task performance, human agency, and human learning. These outcomes underline the baseline proposition of this study that a beneficial augmentation necessitates an explicit consideration of human-AI interaction, and how algorithmic output is designed so that it can match human sensemaking. Figure 1 depicts the conceptual model of this study.

Task Performance. Task performance refers to productivity gains generated through human-AI interactions. Eventually, organizations are interested in AI augmentation to provide some benefits or competitive advantage, e.g., in terms of efficiency, accuracy, or output quality of the task being completed (Gregory et al., 2021; Kemp, 2023; Raisch & Krakowski, 2021). Depending on the nature of the augmented task, different outcomes such as decision accuracy (Fügener et al., 2021), output quality (e.g., Noy & Zhang, 2023) or problem search scope (Raisch & Fomina, 2024) are necessitated.

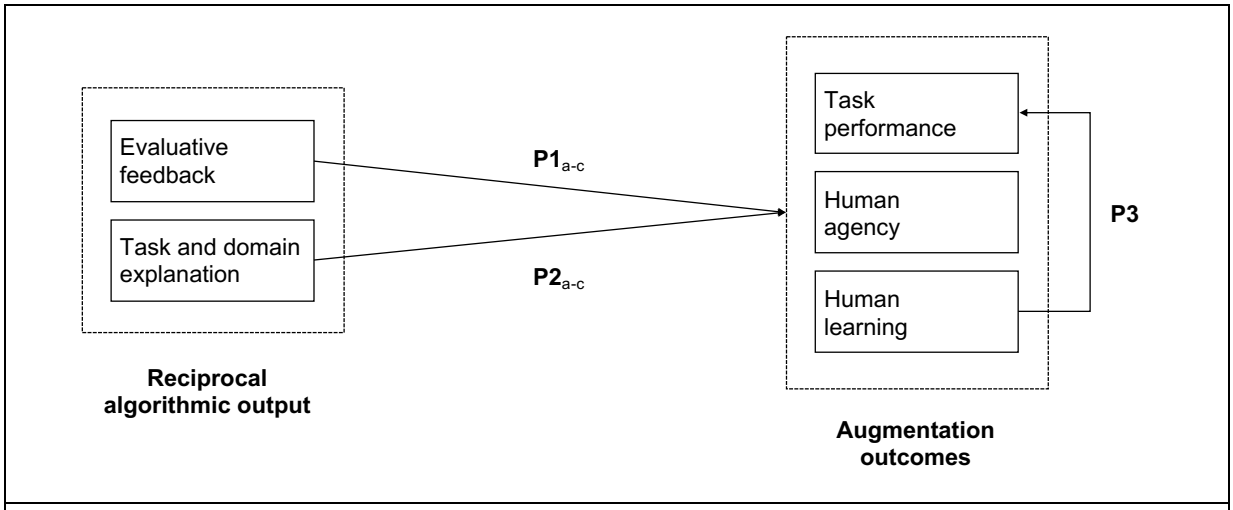

Figure 1. Model for Augmentation in Human-AI Interaction

Human Agency. Human agency refers to a human's capacity to act with intent and free will (Emirbayer & Mische, 1998; Giddens, 1979). Agency is closely correlated with autonomy and control as it enables an agent to influence their direct actions and their environment. From an individual perspective, exercising agency in the workplace has positive direct effects on workers' sense of purpose, motivation, performance, and creativity (e.g., Morgeson et al. 2005; Wu et al., 2015). Human agency is inextricably linked to human-AI interaction: AI is acting, or perceived to be acting, with increased agency (Baird & Maruping, 2021; Murray et al., 2021; Schmitt et al., 2023). Certain designs of human-AI interaction thereby enable workers to forgo their agency or even replace human agency (e.g., Möhlmann et al., 2021). Human-AI interactions that (un)intentionally reduce human agency may, or may not, enable short-term productivity gains, yet at the expense of important individual-level outcomes. Systems becoming increasingly agentic underlines the importance of considering how agency is shared between humans and AI. Extending this line of reasoning, organizations can strengthen human agency by explicitly designing how the output of AI is framed.

Human Learning. This study argues that human learning is core to augmentation. Learning is concerned with a rich depth of information processing that enables the learner to master a skill or topic sustainably (Vygotsky, 1978). Learning is fundamental for workers to develop and maintain skills. If AI is taking over (sub)parts of our tasks, it is to be expected that workers will use their skills in reduced or selected capacity, and, in the long-term, might even lose these skills. This has been shown in early automation studies where human workers take over a supervisory role over a system: delegating a task and using human workers for

the difficult task instances only, workers are prone to lose their task-relevant skills and situational awareness (Billings, 1991; Sheridan, 1987). In a more recent study, Beane (2019) found that dividing tasks between robots and humans could increase efficiency, yet at the expense of workers' learning. Learning is also fundamental to strengthening humans' capabilities to provide feedback to AI-based systems (i.e., reinforcement learning) and to supervise AI-based systems' actions (Zagalsky et al., 2021). This becomes particularly important in light of 'self-consuming' generative models: being increasingly trained with its own generated output, the diversity and novelty of generative models' output decreases (Alemohammad et al., 2023). In short, if there is no human learning, there is unlikely to be AI learning in the future.

Leveraging literature on human learning and reasoning, our model introduces a conceptual model of reciprocal algorithmic output to enable multiple beneficial augmentation outcomes. In the following, we theorize how two key aims of reciprocal design—evaluative feedback, and task-focused explanations—can be leveraged by organizations and system designers.

The Design of Reciprocal Algorithmic Output

This study argues that the design of algorithmic output is a key driver in AI augmentation. To better understand what this study implies by design of output, we refer to Zagalsky et al.'s (2021, p. 3) distinction between the functional and communicational level of human-AI interaction: "*the functional level* […] determines who does what, i.e., the task allocation between human and machine and *the communication level* […] determines what and how is communicated between them." As such, the design of algorithmic output is concerned with the communication level (with implications for task allocation, however).

Reciprocal algorithmic output is a form of algorithmic output that provides the user with evaluative, reflection-provoking input. As opposed to recommendation-based output, reciprocal output does not provide the user with an outcome-focused suggestion the user can simply adopt or reject. As the term *reciprocal* suggests, output builds on a user's initial input or judgement, and can thereby not be made sense of or developed on its own. Vice versa, users cannot draw a conclusive result from the output. A task outcome is hence co-created and requires human input. We can think of algorithmic output moving from prescriptive towards supportive implications for human sensemaking. Reciprocal algorithmic output can be thought of as 1) fostering users' evaluation of possible outcomes by providing feedback and critique on users' initial input, and about 2) increasing users' understanding of the augmented task and related domain. We explain how each of these two aspects address the interaction challenges identified in human-AI interaction. These arguments form the supporting logic for the baseline proposition that reciprocal output increases the potential for beneficial augmentation outcomes (see Figure 1).

Feedback to induce reflection. Evaluative feedback is the phrasing and form of output so that it induces reflection and forces the user to (re)evaluate their initial judgment (Miller, 2023).[1] Evaluative feedback focuses on provoking new thoughts and to think about a task differently. Evaluative feedback may be viewed in contrast to outcome-focused feedback, as part of which users are provided with a recommended best or correct answer. Other than outcome-focused feedback, evaluative feedback requires the user to take action as it is not possible to simply accept or reject the output (see Table 2 for a comparison). This is because evaluative feedback is not framed in a way that is congruent with a task decision or a task outcome.

Evaluative feedback may include open-ended questions such as offering alternative solutions or critiquing humans' input (King, 1990, 1995). It may also involve pro and contra arguments that provide specific feedback to a user's judgement or solution (Miller, 2023). Think of a medical expert receiving pro and contra arguments for their diagnosis without the output making an explicit suggestion of whether the patient has cancer or not. As such, the output is not explicitly confirming or rejecting the expert's judgement. It is up to the expert to decide whether and how to consider these arguments.

Evaluative feedback is an attempt to overcome key limitations in AI-based systems' computational nature, e.g., model brittleness, and resulting implications for human-AI interaction. As machine learning models are largely based on statistical likelihood inferred from selected training data, they also entail an unpredictability towards providing inaccurate and fabricated output (Townsend et al., 2024). As evaluative feedback is open-ended and undetermined in terms of a specific outcome, the algorithmic output, per

[1] The term *evaluative feedback* has been inferred from Miller's (2023) hypothesis-driven framework for explainable AI in decision-making.

default, cannot be incorrect. In turn, users cannot rely or become fixated on an incorrect recommendation. At most, evaluative feedback might not be perceived as very helpful or appropriate, which allows the user to follow their initial judgement or consider other information. In turn, we expect:

Proposition 1a: Task performance does not decrease when receiving evaluative feedback (as compared to outcome-focused feedback).

Per design, evaluative feedback does not allow users to become fixated on the output as they cannot unquestioningly follow the output without their own sensemaking. Evaluative feedback represents a more reciprocal form of output as an understanding is co-created between the user's input and the algorithmic output. As the evaluative feedback is directly tied to and personalized towards the user's input, the algorithmic output cannot stand as an individual, separate piece of information. Evaluative feedback thereby balances against humans' tendency to be easily induced to follow algorithmic output without engaging in too much cognitive effort (Assumption 3, see Logg et al., 2019; Jussupow et al., 2021; Lebovitz et al., 2022). Evaluative feedback thereby also overcomes humans' confirmation bias as the framing of the output makes it incomparable to humans' initial judgement. Returning to the medical example, evaluative feedback is neither congruent with an expert's diagnose of 'disease' or 'no disease'. As such, evaluative feedback necessitates human agency:

Proposition 1b: Human agency increases when receiving evaluative feedback.

The benefits of evaluative feedback are expected to extend to human learning. According to reciprocal learning theory, learning is more effective when output induces critical reasoning and a co-production of knowledge (Jörg, 2004; Vygotsky, 1978; Zagalsky et al., 2021). Diversity in perspectives has also proven beneficial in problem solving contexts (Hong & Page, 2001). Humans' focus on goal completion prevents substantive engagement with output (Assumption 3). Humans' unwillingness to expend cognitive effort thus also prevents learning. However, learning may be enabled through the design of algorithmic output (Carroll & KcKendree, 1987).

Dey et al. (2018) provide a visual design of evaluative feedback for navigation support. They operationalized output as outcome-focused feedback (i.e., map with own location, directional arrows, and recommendation in which direction to turn) versus evaluative feedback (i.e., map with own location only). While both forms of feedback led to effective navigation, the evaluative feedback also led to greater learning.[2] Framing output in a reciprocal manner which necessitates the human to be engaged is therefore expected to increase human learning:

Proposition 1c: Human learning increases when receiving evaluative feedback.

Attributes	Variants	Definition	Examples
Nature of Feedback	Evaluative	Open-ended, reflection-provoking critique	Pro and contra arguments for a human's classification
	Outcome-focused	Closed, conclusive recommendation	Unequivocal classification (e.g., 'disease' / 'no disease')
Focus of Explanation	Task and domain	Improve users' understanding of the task and domain	Linking output to domain knowledge
	System	Improve users' understanding of the underlying model	Feature importance for (classification) model predictions

Table 2. Key Attributes of Reciprocal Algorithmic Output

Explanations to improve task understanding. Improving human understanding of the augmented task attempts to direct explanations of an AI-based system towards information that is relevant to the task. Before diving into the expected effects of task- and domain focused explanations, it is worthwhile to review

[2] Their study also illustrates that evaluative feedback is applicable to different types of tasks and human-AI interaction modalities.

the design of prevalent explanations in human-AI interaction. System-focused explanations address the opacity challenge of AI-based systems by improving users' understanding of the AI-based system (Kemp, 2023). An explanation of how the AI-based system works, e.g., by illustrating the importance of individual input features in a classification prediction model, can provide evidence and justify the algorithmic output the system has arrived at. This is in line with Toulmin's model of argumentation in that explanations can serve as justifications for the offered knowledge (Toulmin, 2003). Explanations about an AI-based system and its inner workings can thus strengthen the persuasiveness of a recommendation by signaling validity and trustworthiness (Dhaliwal & Benbasat, 1996).

System-focused explanations face multiple challenges, though: one, not all models follow a training paradigm and a prediction logic that are retraceable and, hence, possible to explain (Kemp, 2023). Two, it is unclear whether system-focused explanations match human sensemaking, i.e., whether a better understanding of the system enables humans to make better decisions. Empirical findings increasingly converge to the idea that the intended benefits of explanations do not materialize (e.g., Bauer et al., 2023; Schoeffer et al., 2024). Three, previously reviewed interaction challenges raise the question of whether we want to strengthen algorithmic output's persuasiveness. In theory, system-focused explanations can help increase task performance if the algorithmic output is correct (i.e., persuade the user to follow the algorithmic output) and if there is an underreliance (i.e., aversion against the output although it is correct). Considering overreliance, selected engagement, and the unpredictability of AI-based systems' output, explanations that strengthen the persuasiveness and perceived reliability of the system can be harmful and misleading (Lakkaraju & Bastani, 2020).

Explanations can also focus on providing complementary information on a task or domain. Gregor and Benbasat (1999) discussed the benefits of explanations for IS. Next to making a prediction or problem-solving logic traceable, explanations' content can focus on providing 'deep' domain knowledge or definitional and terminological information relevant to the domain or task. In the context of a nutrition classification task, Gajos and Mamykina (2021, p.801) supported humans with task-relevant information, such as "milk is a significant source of carbohydrates". This explanation does not recommend users which answer in the classification task is correct yet offers users to engage with the task-relevant information to make a more informed decision. They found that while both recommendation-based output and explanations improved users' performance, task-focused explanations only also improved users' learning.

Providing additional information on a task and the domain as part of which the task is executed is essential for addressing interaction challenges with AI. Like evaluative feedback, explanations focused on task- and domain-understanding are likely to overcome interaction challenges of overreliance and selected engagement as these types of explanations do not provide justification for a suggested outcome or solution. And if considered, task-focused explanations provide the human with domain knowledge or terminological information that is potentially useful for specific and upcoming task instances. As suggested by Dhaliwal and Benbasat (1996, p.357), "a [system] can only help decision-makers make better judgements if it assists them in learning or better understanding the task environment". We therefore posit:

Proposition 2a: Task performance does not decrease when receiving explanations that focus on the task and the domain (as compared to explanations that focus on the system).

Providing humans with information that is potentially helpful yet is not a cue for a system's accuracy or capabilities necessitates human deliberation, since a human must consider how the complementary information helps their sensemaking. A shallow engagement with the explanation is neither sufficient to be persuaded of some advice nor to make a more informed decision. The role of an explanation thereby shifts from providing evidence to persuade a user towards providing information to enhance a user's understanding. Through this deliberation and gain in domain understanding, humans reduce their dependency on the explanations as a cue for authority and reliability. Accordingly, we posit that:

Proposition 2b: Human agency increases when receiving explanations that focus on the task and the domain.

Improving task understanding reflects the aim of algorithmic output to match human cognitive processes. Increasing human task and domain understanding is essential for learning (Carroll & McKendree, 1987; Vygotsky, 1978). Terminological explanations, for instance, can help with short-term learning, as well as in the context of complex tasks, where definitional information can enable the user to make a decision. We thus predict:

Proposition 2c: Human learning increases when receiving explanations that focus on the task and the domain.

Designing for reversibility. A pragmatic rule that offers further distinction of reciprocal algorithmic output from previous forms of recommendation-based output is that of reversibility (see Table 3 for a comparative overview). Forms of reciprocal output aim to strengthen humans' competencies and enforce cognitive engagement by design, whereas forms of outcome-focused and persuasive advice and their effectiveness assume that humans are cognitively engaged. This distinction implies different (long-term) effects algorithmic output can induce. More specifically, we expect that the effects of reciprocal output persist even when an AI-based system is removed:

Proposition 3: When receiving reciprocal algorithmic output, the impact of AI augmentation persists even after an AI-based system is removed.

This proposition assumes that humans are rationally bound and will attempt to minimize any work-related effort (Assumption 3). If there are no monetary or regulatory incentives for humans, recommendation-based algorithmic output changes task outcomes by inducing humans to adjust their behavior without requiring the human to be cognitively engaged (Carroll & McKendree, 1987; Gregor & Benbasat, 1999). As human thought processes and competencies are left largely unchanged, knowledge will likely deteriorate over time (e.g., Abbas et al., 2024). As a result, performance is likely to revert to the pre-AI introduction state, or even worsen. If reciprocal output requires humans to expend effort, they can enhance their competencies and learn. In addition, these competencies should remain stable as humans do not have the opportunity to 'outsource' their cognitive effort. As a result, the implied beneficial impact of reciprocal output should persist once the AI-based system is removed from the augmentation context.

Dimension	Automation and task delegation	Recommendation-based algorithmic output	Reciprocal algorithmic output
Main intervention target	Task outcome: efficiency and effectiveness	Human behavior: reliance on algorithmic output	Human competencies: learning and agency
Goal	Focused: improve task performance	Focused: improve task performance (and human understanding of AI-based system)	Diverse: ensure human agency, learning, and improve task performance
Intervention	Accuracy: leverage AI capabilities while overcoming human cognitive limitations	Persuasiveness: provide evidence to make human understand and rely on the system	Critique: provide feedback to make human reflect and expend effort on the task
Underlying assumption	AI is systematically superior; human is cognitively bound	Output enables human to be cognitively engaged and to rely on AI in the right instances	Human cognitive engagement is malleable
Normative implications	Might violate human agency and accountability Possibility of false information, inaccuracy, and bias		Necessitates human engagement, control and accountability
Reversibility	Once AI is removed, task performance reverts to pre-AI introduction (or even worsens due to de-skilling)		Implied effects should persist once AI is removed

Table 3. Comparative Overview of Reciprocal Algorithmic Output

Discussion

This paper illuminates a pathway to enabling beneficial human-AI interaction by developing a theory of reciprocal design. Rooted in seminal conceptualizations of augmentation (Engelbart, 1962; Licklider, 1960), the paper argues that desirable and complementary interaction outcomes other than efficiency gains should

be considered in augmentation scenarios. The paper further argues that providing evaluative and thought-provoking feedback, as well as increasing users' task and domain understanding, are two design propositions that organizations and system designers may use to overcome current human-AI interaction challenges. In doing so, the paper offers a model for explaining how and under which conditions humans can benefit from AI augmentation.

Contributions

This work produces several contributions to the management theory, human computer interaction, and behavioral IS literatures. This paper contributes to the research on AI augmentation by suggesting the design of algorithmic output to hold a central role when understanding the effects of AI-based systems in human-AI interactions. There are a myriad of factors shaping the impact of AI-based systems on task augmentation. When AI is introduced in the workplace, there are different potential interaction options, and these should be—yet most often are not—actively considered by organizations and decision-makers.

The reciprocal output design pathway holds the promise of reconciling human-AI interaction challenges as they emerge in the empirical literature and in the real world. The IS literature has identified unintended and unconsidered implications when humans receive algorithmic output, including an overreliance on algorithmic output and a selected engagement with the output that does not help humans to improve but worsen their performance (e.g., Abdel-Karim et al., 2023; Jussupow et al., 2021; Lebovitz et al., 2022). A selected number of papers have considered the use of explanations so that humans can make better sense of the output and the underlying AI-based system. However, these studies find explanations to not improve or even exacerbate interaction challenges (Bauer et al., 2023; Gajos & Mamykina, 2022; Rudin, 2019). We introduce a design perspective and illustrate how prevalent forms and designs of algorithmic output limit humans in their agency and allow humans to become fixated on the output. We also explain how these interaction challenges can be overcome by explicitly considering how algorithmic output is designed. Framing output in an evaluative manner and increasing human understanding of the task offers a desirable path forward for human-AI interaction.

The design pathway also holds the promise to extend theoretical frameworks on AI augmentation. Prior frameworks have described interactions between humans and AI, addressing task- and outcome-level considerations (Fügener et al., 2021; Raisch & Fomina, 2024; Teodorescu et al., 2021) or system capabilities (Murray et al., 2021). Our work offers a micro perspective on specific design requirements for algorithmic output in augmentation scenarios. This study suggests a design instantiation of algorithmic output for different types of augmented tasks. The conceptual model can thus explain how, e.g., Murray et al.'s (2021) concept of augmenting technologies, or Rai et al.'s (2019) concept of task assemblage, could be instantiated on a communication and interaction level. More so, we complement existing frameworks by suggesting that human-AI interaction should enable desirable outcomes beyond efficiency gains.

Doing so, this study also breathes new life into the study of and theoretical frameworks on agency by explicitly treating the preservation of human agency as a central, non-negotiable element to augmentation. This preservation has largely been implicit, or even unconsidered in empirical work. There has been a predominant, almost exclusive, focus on productivity gains attainable through AI augmentation (Murray et al., 2021). We find that in numerous empirical settings, humans' possibility to act with intent and to learn is (unintentionally) reduced by the way algorithmic output is designed. At the same time, existing theoretical frameworks largely focus on the ontological debate of system agency (e.g., Baird & Maruping, 2021; Kemp et al., 2023; Murray et al., 2021). Such an ontological perspective is detrimental to understanding how emerging technologies induce a shift in the locus of agency between human and the technology they are interacting with, and how such locus of agency impacts organizational practices. This paper builds on these ontological arguments of system agency. Adopting a sociotechnical design perspective, we view system agency as modifiable. Under this view, AI-based systems' capabilities can be leveraged to personalize human-AI interactions and to strengthen human agency.

Future Research

This paper offers several avenues for future research. The first and most obvious is the empirical testing of the conceptual model. Reciprocal algorithmic output can be instantiated in different ways. Some of them have been explored in, e.g., crowdsourcing contexts in the form of critical questions (Lekschas et al., 2021)

or navigation contexts in the form of non-suggestive spatial maps (Dey et al., 2018). We view reciprocal output to be applicable to both more creative and open-ended, as well as more structured decision-making tasks. The design's validity and usefulness should thus be investigated for different types of tasks with varying degrees of outcome and solution space multiplicities (Campbell, 1988). Our illustrations of different forms of algorithmic output (e.g., Table 2) provide concrete examples of and guidance on how our proposed model could be examined in an empirical and practical context.

In addition to testing the model, revisiting the boundary conditions of the model's theory building raises questions to be addressed in future research. This study assumes that diverse implications for organizations and their individuals are important to AI augmentation (Assumption 2). Our conceptualization of augmentation outcomes should not suggest that other, e.g., society-level outcomes are irrelevant. For instance, Teodorescu et al. (2021) illustrate how AI augmentation can achieve fairness. Beyond testing the validity of our conceptual model, future research could explore how to extend our model to other important outcomes or how other outcomes mediate our hypothesized effects. In addition, if an organization or individual strives to maximize one augmentation outcome only (and neglect the others), other designs of algorithmic output might be more appropriate. For decision-making, selectively providing outcome-focused feedback can be effective in improving task performance, for instance (Fügener et al., 2021).

More so, our conceptual model treats organizational boundaries as an underlying assumption of how AI-based systems are deployed and enacted upon by individual workers (Assumption 3). Thinking more explicitly about organizational boundaries could present an important moderator of how algorithmic output affects learning, agency, and task performance—even to the extreme that expected benefits could disappear. For instance, individuals might work under so much time pressure that they do not have the time to process the output of an AI-based system at all, and the benefits of feedback and complementary task information become obsolete. In a similar vein, managerial expectations of whether and how workers should leverage AI-based system could act as a form of organizational control, thereby implicitly pressuring workers into following the algorithmic output regardless of their own sensemaking (e.g., see Bannon, 2023).

Another important boundary condition that deserves further attention is that of individual differences, including human expertise. It is to be expected that learning effects for highly skilled and experienced individuals is more limited than for novices (Berlin & Jeffries, 1992). Human expertise could present a moderating effect of our main causal relationship in that algorithmic output is expected to have a reduced positive impact on task performance and learning for more experienced individuals. Vice versa, if a human is lacking knowledge and experience, output and related explanations allow for more learning and task improvement. We expect our reversibility effect (Proposition 3) to hold regardless of expertise level though, as skill maintenance is relevant to all skill levels.

Some tasks are rightfully suited to be automated, e.g., dangerous or low-impact and mundane work (Walsh & Strano, 2018). Certain tasks offer to be fully automated, e.g., when we do not care about matters of accountability or liability, and we are solely concerned with task completion or productivity gains as a task outcome. The conceptual model of this study does not attempt to motivate that all human skills for every existing task should be retained and augmented. We make an argument for tasks that are currently considered suitable or even necessary to be augmented (Assumption 1). An important future research avenue is therefore to gain a better understanding of the types of tasks that are crucial to augment, as well as what new tasks and skills the use of AI-based systems could generate (Acemoglu et al., 2023).

Conclusion

As AI will play a more and more dominant role in organizations, a key strategic question will be how organizations and their workers can benefit from AI beyond an excessive focus on performance gains. This paper takes a first step toward addressing that question by proposing a sociotechnical design perspective to overcome current challenges in human-AI interaction and to leverage the interactive capabilities of contemporary AI-based systems. This work leads to a better understanding of how algorithmic output can be designed to foster human-AI interaction without forgoing human agency. Such interactions promise to realize performance gains while allowing for further desirable augmentation outcomes, including human agency and learning.

References

Abbas, M., Jam, F. A., & Khan, T. I. (2024). Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students. *International Journal of Educational Technology in Higher Education,* 21(1), 10.

Abdel-Karim, B. M., Pfeuffer, N., Carl, K. V., & Hinz, O. (2023). How AI-Based Systems Can Induce Reflections: The Case Of AI-Augmented Diagnostic Work. *MIS Quarterly,* 47(4).

Acemoglu, D., Autor, D., & Johnson, S. (2023). Can We Have Pro-Worker AI?" *CEPR Policy Insight 123*.

Acemoglu, D., & Johnson, D. (2023). *Power and progress: our thousand-year struggle over technology and prosperity*, *PublicAffairs*.

Acemoglu, D., & Restrepo, P. (2022). Tasks, automation, and the rise in US wage inequality. *Econometrica*, 90(5), 1973-2016.

Alemohammad, S., Casco-Rodriguez, J., Luzi, L., Humayun, A. I., Babaei, H., LeJeune, D., ... & Baraniuk, R. G. 2023. Self-consuming generative models go mad. arXiv preprint arXiv:2307.01850.

Alvesson, M., & Spicer, A. (2012). A stupidity-based theory of organizations. *Journal of Management Studies*, 49(7), 1194-1220.

Arya, V., Bellamy, R. K., Chen, P. Y., Dhurandhar, A., Hind, M., Hoffman, S. C., Houde, S., Liao, Q. V., Luss, R., Mojsilović, A., Mourad, S., ... & Zhang, Y. (2019). One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques. *arXiv* 1909.03012v2.

Baird, A., & Maruping, L. M. (2021). The Next Generation of Research on IS Use: A Theoretical Framework of Delegation to and from Agentic IS Artifacts. *MIS Quarterly*, 45(1).

Bannon, L. (2023). When AI Overrules the Nurses Caring for You. *Wall Street Journal*, https://www.wsj.com/articles/ai-medical-diagnosis-nurses-f881b0fe.

Bauer, K., von Zahn, M., & Hinz, O. (2023). Expl (AI) ned: The impact of explainable artificial intelligence on users' information processing. *Information Systems Research*, 34(4), 1582-1602.

Beane, M. (2019). Shadow learning: Building robotic surgical skill when approved means fail. *Administrative Science Quarterly*, 64(1), 87-123.

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly*, 45(3).

Berlin, L. M., & Jeffries, R. (1992). Consultants and apprentices: observations about learning and collaborative problem solving. *Proceedings of the 1992 ACM conference on Computer-supported cooperative work*, 130-137.

Billings, C.E. (1991). Human-centered aircraft automation: A concept and guidelines. *NASA*, https://ntrs.nasa.gov/search.jsp?R=19910022821.

Carroll, J. M., & McKendree, J. (1987). Interface Advice Issues for Advice-giving Expert Systems. *Communications of the ACM*, 30, 14-31.

Dey, S., Karahalios, K., & Fu, W.T. (2018). Getting There and Beyond: Incidental Learning of Spatial Knowledge with Turn-by-Turn Directions and Location Updates in Navigation Interfaces. *Proceedings of the Symposium on Spatial User Interaction* (SUI '18).

Dhaliwal, J. S., & Benbasat, I. (1996). The use and effects of knowledge-based system explanations: theoretical foundations and a framework for empirical evaluation. *Information Systems Research*, 7(3), 342-362.

Emirbayer, M., & Mische, A. (1998). What is agency?. *American Journal of Sociology*, 103(4), 962-1023.

Engelbart, Douglas C. (1962). Augmenting Human Intellect: A Conceptual Framework. *Ideas That Created the Future*. https://doi.org/10.7551/MITPRESS/12274.003.0024

Fisher, G., Mayer, K., & Morris, S. (2021). From the editors—Phenomenon-based theorizing. *Academy of Management Review*, 46(4), 631-639.

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2021). Will humans-in-the-loop become borgs? Merits and pitfalls of working with AI. *MIS Quarterly*, 45.

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

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation?. *Technological forecasting and social change,* 114, 254-280.

Gajos, K.Z., & Mamykina, L. (2022). Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning. *International Conference on Intelligent User Interfaces* (IUI '22).

Giddens, A. (1984). *The constitution of society: Outline of the theory of structuration*, Univ of California Press.
Gregor, S., & Benbasat, I. (1999). Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS Quarterly*, 23(4), 497-530.
Gregory, R. W., Henfridsson, O., Kaganer, E., & Kyriakou, H. (2021). The role of artificial intelligence and data network effects for creating user value. *Academy of Management Review*, 46(3), 534-551.
Hackman, J.R., & Oldham, G.R. (1976). Motivation through the design of work: Test of a theory. *Organizational Behavior and Human Performance*, 16(2), 250–279.
Hong, L. & Page, S.E. (2001). Problem Solving by Heterogenous Agents. *Journal of Economic Theory*, 97, 123-163
Jörg, T. (2004). A theory of reciprocal learning in dyads. *Cognitive systems*, 6(2).
Jussupow, E., Spohrer, K., Heinzl, A., & Gawlitza, J. (2021). Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence. *Information Systems Research*, 32(3), 713-735.
Kane, G. C., Young, A. G., Majchrzak, A., & Ransbotham, S. (2021). Avoiding an oppressive future of machine learning: A design theory for emancipatory assistants. *MIS Quarterly*, 45(1), 371-396.
Kemp, A. (2023). Competitive Advantage through Artificial Intelligence: Toward a Theory of Situated AI. *Academy of Management Review*.
King, A. (1990). Enhancing peer interaction and learning in the classroom through reciprocal questioning. *American Educational Research Journal*, 27(4), 664-687.
King, A. (1995). Designing the instructional process to enhance critical thinking across the curriculum: Inquiring minds really do want to know: Using questioning to teach critical thinking. *Teaching of Psychology*, 22(1), 13-17.
Lakkaraju, H., Bastani, O. (2020). "How do I fool you?": Manipulating User Trust via Misleading Black Box Explanations. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 79–85.
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. *Organization Science*, 33(1), 126-148.
Lekschas, F., Ampanavos, S., Siangliulue, P., Pfister, H., Gajos, K.Z. (2021). Ask Me or Tell Me? Enhancing the Effectiveness of Crowdsourced Design Feedback. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.
Licklider, J. C. R. (1960). Man-Computer Symbiosis." *IRE Transactions on Human Factors in Electronics* HFE-1 (1), 4–11.
Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151, 90-103.
Miller, T. (2023). Explainable AI is dead, long live explainable AI! Hypothesis-driven decision support using evaluative AI. *ACM Conference on Fairness, Accountability, and Transparency*, 333-342.
Möhlmann, M., Zalmanson, L., Henfridsson, O., & Gregory, R. W. (2021). Algorithmic Management of Work on Online Labor Platforms: When Matching Meets Control. *MIS Quarterly*, 45(4).
Morgeson, F. P., Delaney-Klinger, K., & Hemingway, M. A. (2005). The Importance of Job Autonomy, Cognitive Ability, and Job-Related Skill for Predicting Role Breadth and Job Performance. *Journal of Applied Psychology*, 90(2), 399–406.
Murray, A., Rhymer, J. E. N., & Sirmon, D. G. (2021). Humans and technology: Forms of conjoined agency in organizations. *Academy of Management Review*, 46(3), 552-571.
Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654), 187-192.
Parker, S. K., & Grote, G. (2022). Automation, algorithms, and beyond: Why work design matters more than ever in a digital world. *Applied Psychology*, 71(4), 1171-1204.
Rai, A., Constantinides, P., & Sarker, S. (2019). Next generation digital platforms: toward human-AI hybrids. *MIS Quarterly*, 43(1), iii-ix.
Raisch, S., & Fomina, K. (2023). Combining human and artificial intelligence: Hybrid problem-solving in organizations. *Academy of Management Review*.
Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation–augmentation paradox. *Academy of Management Review,* 46(1), 192-210.
Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence,* 1(5), 206–215.

Schmitt, A.; Zierau, N.; Janson, A.; Leimeister, J. M. (2023). The Role of AI-Based Artifacts' Voice Capabilities for Agency Attribution. *Journal of the Association for Information Systems,* 24(4), 980-1004.

Schoeffer, J., De-Arteaga, M., & Kühl, N. (2024). Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making. *Proceedings of the CHI Conference on Human Factors in Computing Systems* (CHI '24).

Sheridan, T.B. (1987). Supervisory control. In *Handbook of human factors,* G. Salvendy (ed.), Oxford, UK: John Wiley & Sons, 1243–1268.

Te'eni, D. (2001). A Cognitive-Affective Model of Organizational Communication for Designing IT. *MIS Quarterly*, 25(2), 251-312

Teodorescu, M. H., Morse, L., Awwad, Y., & Kane, G. C. (2021). Failures of Fairness in Automation Require a Deeper Understanding of Human-ML Augmentation. *MIS Quarterly*, 45(3).

Toulmin, S.E. (2003). *The uses of argument*, Cambridge University Press.

Townsend, D. M., Hunt, R. A., Rady, J., Manocha, P., & Jin, J. H. (2024). Are the Futures Computable? Knightian Uncertainty and Artificial Intelligence. *Academy of Management Review*.

Vygotsky, L. S. (1978). *Mind in society: Development of higher psychological processes,* Harvard University Press.

Walsh, S. M., & Strano, M. S. (2018). *Robotic systems and autonomous platforms: Advances in materials and manufacturing,* Woodhead Publishing.

Wu, C. H., Griffin, M. A., & Parker, S. K. (2015). Developing agency through good work: Longitudinal effects of job autonomy and skill utilization on locus of control. *Journal of Vocational Behavior*, 89, 102-108.

Zagalsky, A., Te'eni, D., Yahav, I., Schwartz, D. G., Silverman, G., Cohen, D., ... & Lewinsky, D. (2021). The design of reciprocal learning between human and artificial intelligence. *Proceedings of the ACM on Human-Computer Interaction (CSCW2),* 1-36.

Zuboff, S. (1985). Automate/informate: The two faces of intelligent technology. *Organizational Dynamics,* 14, 5-18.