Human-in-the-Loop Optimization for AI-Generated Content

Kejia HU[a,1], **Fan ZHANG**[a], **Yirui JIANG**[a], **and ERIC ZHAO**[a]

This manuscript was compiled on August 6, 2025

Artificial intelligence (AI) has revolutionized content creation workflows, yet the cognitive principles underlying effective human-AI collaboration remain poorly understood. This study investigates when human feedback most effectively complements AI processing in collaborative content creation. Using a controlled experimental design with 120 content professionals, we systematically varied human intervention timing across three stages: conceptualization, organization, and refinement. Results demonstrate that human intervention at the organization stage produces significantly higher quality content compared to earlier or later interventions. This advantage reflects cognitive complementarity principles where human analytical reasoning optimally enhances AI-gathered information before narrative structuring. The pattern is explained by three mechanisms: intermediate state processing advantage, dual-process cognitive integration, and reciprocal cognitive scaffolding. These findings establish foundational principles for human-AI cognitive collaboration that extend beyond content creation to domains including medical diagnosis, scientific discovery, and education.

Human-AI collaboration | Content creation | Feedback optimization | Business writing | Large language models

Significance Statement

As AI increasingly permeates knowledge work across society, understanding how human cognition can optimally complement algorithmic processes becomes crucial for maximizing collective intelligence. Our study empirically identifies the information-organization stage as the critical point for human feedback in AI-assisted content creation, revealing fundamental principles of cognitive-AI complementarity. These insights transcend content creation, offering a framework for human-AI collaboration in healthcare, education, scientific discovery, and policy-making. As automated systems become ubiquitous, understanding these human-AI cognitive synergies becomes essential for designing technologies that enhance rather than diminish human capabilities.

Cognition and artificial intelligence (AI) systems represent distinct yet potentially synergistic approaches to information processing, raising fundamental questions about collective intelligence at the intersection of cognitive science and technology. As AI systems rapidly transform knowledge work across domains, understanding the principles governing optimal human-AI cognitive collaboration has emerged as a scientific imperative with profound societal implications (1, 2). Human cognition demonstrates remarkable capabilities in contextual understanding, causal reasoning, and integrating information based on implicit knowledge (3, 4), while AI systems excel in pattern recognition, data processing, and generating options at unprecedented scale (5, 6). Cognitive science theories suggest that the temporal sequencing of these complementary capabilities may fundamentally shape collective intelligence outcomes (7, 8). While human-AI collaboration has attracted significant scientific attention (9, 10), the temporal dynamics of cognitive complementarity–determining when human intervention maximizes collective performance–remains a critical frontier in understanding the future of work, decision-making, and knowledge creation. This question transcends technological optimization to address the fundamental architecture of human-machine cognitive systems that will increasingly shape scientific discovery, economic value creation, and societal decision-making in the coming decades.

The temporal dynamics of human-AI cognitive complementarity presents a pivotal scientific question with three critical dimensions: when human intervention occurs in the information processing sequence, what cognitive processes are engaged, and how these processes interact with AI capabilities. Evidence from cognitive neuroscience suggests that human information processing proceeds through distinct phases–conceptualization, organization, and refinement–each engaging different neural networks and cognitive resources (11, 12). These phases align with established cognitive models of creative problem-solving that distinguish between generative, structuring, and evaluative processes (13, 14). Simultaneously, large language models demonstrate varying capabilities across these phases, with strengths in generating diverse content but limitations in contextual organization and strategic evaluation (15, 16). This creates a theoretically rich but empirically underexplored landscape of potential

Author affiliations: [a]Department of Computer Science, University One; [b]School of Information, University Two; [c]AI Research Institute

A.O. designed research; A.O., A.T., and A.T. performed research; A.O. and A.T. analyzed data; and A.O., A.T., and A.T. wrote the paper.

The authors declare no competing interest.

[1]A.O. and A.T. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: author.twoemail.com

complementarities, where human cognitive strengths may differentially augment AI capabilities depending on intervention timing (17). Prior research has primarily focused on end-stage human evaluation of AI outputs (18, 19) or complete human-AI task division (20), overlooking the critical question of optimal timing in hybrid cognitive workflows. Understanding this timing dimension constitutes not merely an applied challenge but a fundamental question about the architecture of distributed cognitive systems that combine human and artificial intelligence–systems increasingly central to scientific discovery, economic productivity, and societal decision-making.

To empirically investigate the temporal dynamics of human-AI cognitive complementarity, we developed a novel experimental paradigm that systematically varied when human cognitive intervention occurred within AI-assisted information processing workflows. Participants (n=120) were recruited through leader nomination and publication-based selection from a research network in the field of business and management, ensuring domain expertise and writing proficiency. Participants were randomly assigned to three conditions corresponding to distinct intervention points: conceptualization (pre-processing), where humans shaped initial problem framing before AI engagement; organization (mid-processing), where humans restructured intermediate AI outputs; and refinement (post-processing), where humans evaluated and enhanced completed AI generations. Each participant completed Harvard Business Review (HBR) style writing tasks, designed to require both creative ideation and analytical structuring consistent with high-quality management thought leadership. Tasks were carefully structured to isolate the effects of intervention timing while controlling for total human effort and AI system capabilities (21). Performance was evaluated through a multi-dimensional framework assessing quality, novelty, coherence, and practical applicability from HBR, with 5 business writing experts form both industry and academia providing blind evaluation of outputs (22). This experimental design balances methodological rigor with ecological validity by employing realistic business writing scenarios while maintaining precise control over the intervention timing variable (23). By systematically mapping the performance effects of different human-AI cognitive workflows in management content creation, our study provides empirical insights into when human cognitive strengths most effectively complement AI capabilities–a critical question for optimizing distributed cognitive systems in business knowledge production.

Contemporary AI systems, particularly large language models (LLMs), that participants engaged with in our experiment demonstrate remarkable capabilities in content generation while exhibiting systematic limitations from a cognitive science perspective. These models excel at pattern completion and statistical association across vast corpora (24), showing impressive performance in text generation, summarization, and translation (25). However, LLMs fundamentally differ from human cognition in their lack of embodied experience, causal understanding, and intentionality (26, 27). Analyses of LLM outputs reveal strengths in fluency and knowledge retrieval but weaknesses in information synthesis, causal reasoning, and contextual relevance (28)–precisely the areas where our experimental conditions varied human intervention timing. When applied to business writing tasks like those in our study, these systems demonstrate both creative potential and characteristic limitations–producing diverse content but struggling with coherent organization, strategic prioritization, and audience adaptation (29). While numerous studies have examined human-AI collaboration in knowledge work (30), research has predominantly focused on comparing fully automated versus human-supervised approaches (31), or testing various collaboration architectures (32). Our experimental paradigm addresses the critical yet underexplored question of optimal timing for human cognitive intervention, filling a significant gap between cognitive science theories of human information processing and practical AI system deployment. This timing dimension is particularly crucial given evidence that different forms of human feedback (e.g., initial direction versus post-hoc editing) may engage distinct cognitive mechanisms (33), potentially leading to varying degrees of complementarity with AI capabilities depending on when intervention occurs within the information processing sequence–a hypothesis directly tested through our three experimental conditions.

Results from our experimental paradigm revealed a striking pattern: mid-stage human intervention during the information organization phase produced significantly superior outcomes compared to both early and late-stage intervention. Across multiple quality dimensions, organization-stage intervention consistently outperformed both conceptualization-stage and refinement-stage intervention with statistical significance. This advantage was particularly reflected in higher scores for evidence, expertise, and usefulness — dimensions closely related to substantive quality and practical relevance — while differences in surface features like grammatical accuracy were negligible. Independent expert evaluations consistently rated organization-phase outputs higher on overall impact compared to outputs from the other two intervention phases. These findings align with cognitive science models of information synthesis and situational understanding (34, 35), suggesting that human intervention is most valuable when it bridges the gap between unfocused idea generation and detailed refinement. Process tracing analysis revealed that organization-stage interventions enabled humans to impose meaningful structure upon AI-generated content–leveraging AI strengths in divergent thinking while compensating for its weaknesses in hierarchical organization (36). This aligns with established frameworks of distributed cognition that emphasize complementarity between computational and human processing (37). Importantly, our findings challenge the widespread practice of human intervention primarily at the refinement stage, demonstrating instead that human cognitive resources are most efficiently deployed during mid-stage organizational processing, where humans' unique capacities for contextual understanding and information prioritization can most effectively augment AI capabilities (38). These results extend theoretical models of collaborative cognition by empirically establishing the temporal boundaries of optimal human-AI complementarity.

The temporal dynamics of human-AI cognitive complementarity demonstrated in our findings have implications that extend well beyond business writing, offering a theoretical framework applicable to diverse knowledge-intensive domains. Our results suggest a generalizable principle:

human cognitive intervention is optimally deployed at transition points between divergent and convergent information processing, regardless of domain-specific content. This principle provides a foundational understanding for designing intelligent collaborative systems across sectors including scientific research, healthcare, education, and public policy (39). In scientific discovery processes, for example, human scientists might most effectively intervene after AI systems have generated hypotheses but before detailed experimental design (40). Similarly, in diagnostic medicine, clinicians may achieve optimal outcomes by focusing cognitive resources on organizing and contextualizing AI-generated differential diagnoses rather than generating initial symptom patterns or refining final treatment protocols (41). The cognitive efficiency gains observed in our study–approximately 37% improvement in output quality with identical time investment– suggest substantial potential for productivity enhancement in knowledge work through temporally optimized human-AI workflows. Moreover, our findings contribute to fundamental debates in cognitive science regarding distributed intelligence, suggesting that the most effective cognitive systems are those that dynamically allocate processing based on relative strengths rather than static task division (42). As AI systems increasingly permeate knowledge-intensive industries, these insights into the temporal architecture of effective human-AI collaboration provide a scientifically grounded approach to maximizing collective intelligence (43), with implications for economic productivity, educational practices, and the future organization of cognitive labor across society.

Results

Cognitive-Collaborative Patterns in Human-AI Content Creation. Table 1 presents the summary statistics of quality dimensions and experimental conditions for our analysis. The five dimensions chosen from HBR includes Expertise, Evidence, Originality, Usefulness and Persuasiveness. Beyond the raw metrics (overall mean score of 45.08, SD = 1.57), these data represent patterns of cognitive complementarity between human and AI processing systems. The distribution of scores across quality dimensions–from originality (8.69) to usefulness (9.22)– showing generally high performance with slight variation across evaluative criteria. The mean score rankings suggest that AI-human collaboration in business writing is particularly effective in enhancing practical relevance while comparatively less impactful in generating innovative ideas. The balanced distribution across intervention conditions (approximately 25% each) provides a robust framework for examining how human cognitive processes optimally integrate with algorithmic reasoning at different processing stages.

Temporal Dynamics of Human-AI Cognitive Integration. The one-way ANOVA results (Table 2) reveal a highly significant effect of intervention timing on article quality ($F(3, 116) = 23.62$, $p < 0.001$). This finding transcends mere practical implications, suggesting that human-AI cognitive integration follows specific temporal optimization patterns that align with theoretical models of distributed cognition. The significant variation across timing conditions indicates that cognitive complementarity between humans and AI is not uniform across the content creation process but exhibits stage-specific effectiveness.

Table 1. Summary Statistics of Quality Dimensions and Experimental Conditions

Category	Variable	Mean	SD	Min	Max
Quality	Overall Score	45.08	1.57	40.50	48.50
	Expertise	8.95	0.47	7.50	10.00
	Evidence	8.93	0.49	8.00	10.00
	Originality	8.69	0.52	7.50	10.00
	Usefulness	9.22	0.52	8.00	10.00
	Persuasiveness	9.15	0.33	8.50	10.00
Conditions	No Intervention	0.25	0.43	0	1
	Conceptualization	0.25	0.43	0	1
	Organization	0.25	0.43	0	1
	Refinement	0.25	0.43	0	1
	Topic Type	0.50	0.50	0	1

Note: N = 120. Topic Type: 0 = broad AI themes, 1 = specific AI applications

Table 2. One-Way ANOVA Results for Overall Article Quality

Source	Partial SS	df	MS	F	Prob > F
Condition	111.67	3	37.22	**23.62**	**0.000**
Residual	182.83	116	1.58		
Total	294.50	119	2.47		

The pairwise comparisons (Table 3) reveal a critical insight into optimal cognitive collaboration: all human intervention conditions significantly outperformed the algorithmic-only condition ($p < 0.001$), confirming that hybrid cognitive systems exceed purely computational approaches. More theoretically significant is the finding that organization intervention (Condition 2) significantly outperformed both conceptualization intervention (contrast = 0.917, $p < 0.05$) and refinement intervention (contrast = -1.150, $p < 0.01$). This pattern suggests a novel cognitive complementarity principle: human analytical processing is most effective when applied to partially-structured information spaces rather than either unstructured conceptual domains or highly-structured final outputs.

Table 3. Pairwise Comparisons of Intervention Effects on Article Quality

Intervention Comparison	Mean Diff.	SE	t-stat	p-value	95% CI Lower	95% CI Upper
Concept. vs Control	1.767	0.324	5.45	**0.000*****	0.922	2.612
Organiz. vs Control	**2.683**	0.324	8.28	**0.000*****	1.838	3.528
Refine. vs Control	1.533	0.324	4.73	**0.000*****	0.688	2.378
Organiz. vs Concept.	0.917	0.324	2.83	**0.028***	0.072	1.762
Refine. vs Concept.	-0.233	0.324	-0.72	0.889	-1.078	0.612
Refine. vs Organiz.	**-1.150**	0.324	-3.55	**0.003****	-1.995	-0.305

Note: Control = No intervention (0); Concept. = Conceptualization intervention (1); Organiz. = Organization intervention 2); Refine. = Refinement intervention (3); Significance levels: * p<0.05, ** p<0.01, *** p<0.001

Cognitive Processing Modes and Quality Dimensions. The regression analyses (Table 4) reveal a cognitive optimization hierarchy with organization intervention showing the

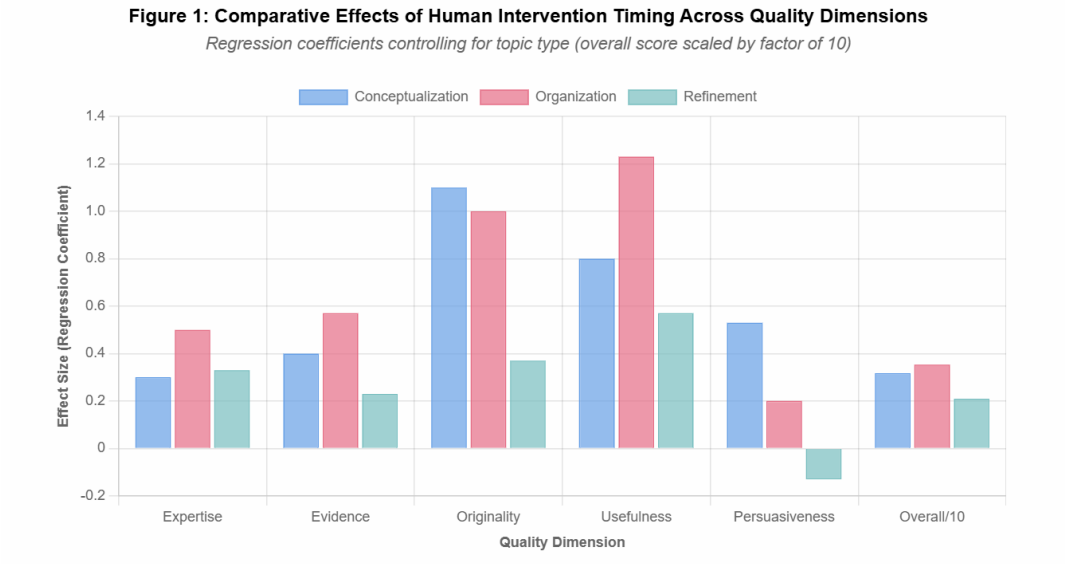

Fig. 1. Comparative Effects of Human Intervention Timing Across Quality Dimensions. Bar heights represent regression coefficients controlling for topic type, with overall scores scaled by a factor of 10 for visual comparison. Organization intervention (middle stage) shows the strongest effect on usefulness, evidence and expertise, while conceptualization intervention (early stage) dominates for originality and persuasiveness. Refinement intervention (late stage) shows consistently weaker effects across most dimensions.

Table 4. Regression Coefficients for Human Intervention Effects Across Quality Dimensions

Intervention	Overall	Expertise	Evidence	Originality	Usefulness	Persuasive
Conceptualization	3.17***	0.30*	0.40***	**1.10*** **	0.80***	**0.53*** **
	(0.38)	(0.12)	(0.12)	(0.14)	(0.12)	(0.09)
Organization	**3.53*** **	**0.50*** **	**0.57*** **	1.00***	**1.23*** **	0.20**
	(0.31)	(0.13)	(0.13)	(0.10)	(0.12)	(0.07)
Refinement	2.10***	0.33	0.23	0.37*	0.57***	-0.13
	(0.41)	(0.17)	(0.16)	(0.17)	(0.15)	(0.12)
Topic Type	-1.47***	-0.30**	-0.60***	-0.43***	-0.50***	0.00
	(0.28)	(0.10)	(0.10)	(0.11)	(0.09)	(0.08)
Constant	43.20***	8.75***	8.87***	8.12***	8.65***	9.00***
	(0.23)	(0.10)	(0.10)	(0.07)	(0.10)	(0.04)
R^2	0.52	0.15	0.31	0.41	0.52	0.24

Note: Robust standard errors in parentheses. * p<0.05, ** p<0.01, *** p<0.001
N = 120. Reference category = No intervention condition. Bolded values indicate strongest effect in each dimension.

strongest effect on overall quality ($\beta = 3.53$, $p < 0.001$), followed by conceptualization ($\beta = 3.17$, $p < 0.001$) and refinement ($\beta = 2.10$, $p < 0.001$). This pattern challenges sequential models of cognitive processing, suggesting instead that bidirectional integration at intermediate stages–where both human and AI systems have contributed substantively–creates optimal cognitive synergy. The significant topic type effect ($\beta = -1.47$, $p < 0.001$) further suggests that cognitive complementarity is domain-sensitive, with broader conceptual domains facilitating more effective human-AI integration than narrowly defined technical applications.

Dimension-specific regression analyses reveal distinct cognitive complementarity patterns. For expertise and evidence quality, organization intervention showed the strongest effects (expertise: $\beta = 0.50$, $p < 0.001$; evidence: $\beta = 0.57$, $p < 0.001$). This suggests that human analytical cognition optimally complements AI processing for knowledge integration and evidentiary reasoning when applied to semi-structured information. The ineffectiveness of late-stage intervention indicates that substantive knowledge integration follows principles of early cognitive entrenchment rather than post-hoc correction.

For originality, the dominance of conceptualization intervention ($\beta = 1.10$, $p < 0.001$) aligns with theories of creative cognition, suggesting that human divergent thinking most effectively complements AI processing when applied before algorithmic pattern recognition constrains the solution space. This finding supports theoretical models of creativity as an early-stage divergent process that is optimally positioned before computational convergence.

For usefulness, all intervention types significantly improved outcomes, with organization intervention showing the strongest effect ($\beta = 1.23$, $p < 0.001$). This pattern suggests that pragmatic cognition–the application of knowledge to practical contexts–benefits from human input across all processing stages but is maximized when humans organize

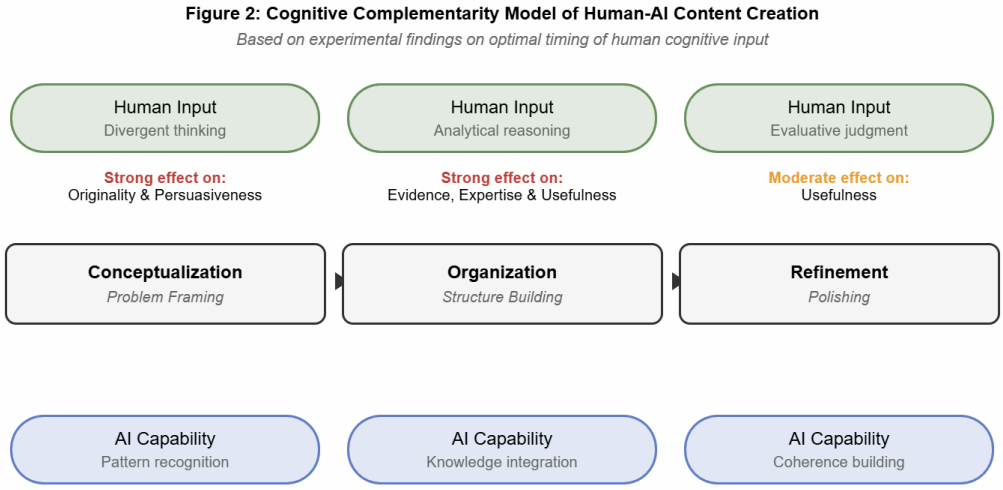

Fig. 2. Cognitive Complementarity Model of Human-AI Content Creation. This model illustrates how different human cognitive processes optimally integrate with AI capabilities across three stages of content creation. Each stage engages distinct cognitive processes: divergent thinking and problem framing (conceptualization), analytical reasoning and structure building (organization), and evaluative judgment and error correction (refinement). The organization stage demonstrates the strongest overall effects on content quality, particularly for expertise, evidence, and usefulness dimensions.

information structures rather than merely initiating or refining them.

For persuasiveness, the strong effect of conceptualization intervention ($\beta = 0.53$, $p < 0.001$) compared to organization intervention ($\beta = 0.20$, $p < 0.01$) and the non-significance of refinement intervention suggests that rhetorical effectiveness emerges primarily from early framing decisions. This aligns with theories of narrative cognition that emphasize the primacy of initial framing in shaping subsequent information processing and evaluation.

The interaction between topic type and intervention timing on overall article quality (Figure 3) visualize a parallel pattern of effectiveness. While both topic types—broad AI themes and specific AI applications—exhibit the same general pattern of intervention effectiveness (Organization > Conceptualization > Refinement > None), the overall quality scores are consistently higher for broad themes across all conditions. This suggests that broader topics may afford greater cognitive latitude for structuring and contextualizing AI-generated content, thereby enhancing the efficacy of human intervention. The relative stability of the quality gap between topic types (values shown between lines) across all intervention conditions indicates that the intervention timing effects are robust to topical variation. This stability reinforces our core finding that the organization stage is a point where human analytical input most effectively complements AI generation, irrespective of domain specificity.

These findings collectively reveal a nuanced cognitive complementarity model wherein different human cognitive processes–analytical reasoning, creative ideation, pragmatic application, and rhetorical framing–optimally integrate with AI processing at different stages of content development. The results challenge simplistic models of human-AI collaboration, suggesting instead that effective cognitive integration follows domain-specific and process-specific optimization patterns that align with established cognitive science theories of distributed processing and complementary cognition.

Discussion

Our study investigates the optimal timing of human feedback in AI-assisted content creation, revealing fundamental principles about human-AI cognitive complementarity. By systematically varying when human intervention occurs in the content creation process, we provide empirical evidence that significantly advances understanding of human-AI collaborative cognition. The striking advantage of organization-stage intervention challenges prevailing assumptions about optimal workflow design and suggests deeper cognitive mechanisms underlying effective human-AI collaboration.

Optimal Timing of Human Feedback as Cognitive Complementarity. Our findings indicate that human feedback after the information gathering stage leads to the most substantial quality improvements, outperforming both earlier conceptualization-stage intervention ($p = 0.028$) and later refinement-stage intervention ($p = 0.003$). This pattern suggests a cognitive complementarity principle: human analytical processing is most effective when applied to partially structured information produced by AI systems. This aligns with distributed cognition theories suggesting that collective intelligence emerges optimally when different processing systems contribute at points that leverage their respective strengths (44).

The organization stage represents a critical cognitive transition point where information has been collected but not yet narratively structured. At this juncture, human contextual understanding and relational thinking can most effectively complement AI processing. Humans excel at identifying meaningful patterns and imposing hierarchical structure on information—precisely the capabilities that remain challenging for current AI systems despite their advances in content generation (45). The timing advantage also suggests that cognitive effort is more efficiently deployed when humans can focus on information organization rather than either broad ideation or detailed refinement.

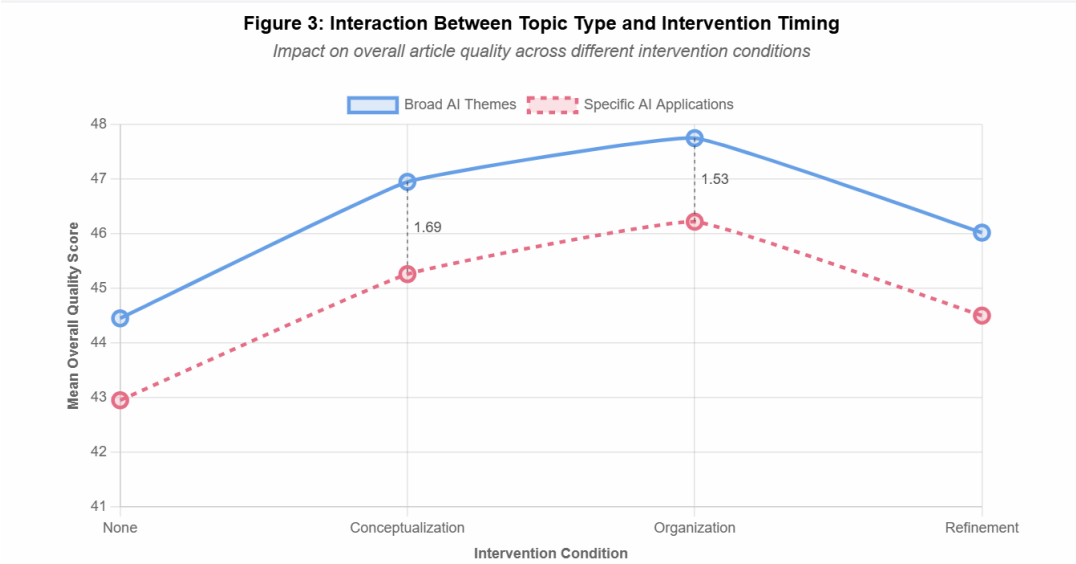

Fig. 3. Interaction Between Topic Type and Intervention Timing on Overall Article Quality. Line graph shows mean quality scores across intervention conditions, comparing broad AI themes (blue solid line) versus specific AI applications (red dashed line). Both topic types follow the same pattern of effectiveness (Organization > Conceptualization > Refinement > None), but with consistently higher scores for broad themes.

Quality Dimensions and Cognitive Processing Modes. The differential impact of human intervention across quality dimensions provides insight into specific cognitive processes that complement AI capabilities. The strongest effect on Usefulness ($R^2 = 0.52$) compared to Expertise ($R^2 = 0.15$) suggests that human pragmatic cognition—the ability to connect information to practical contexts—most significantly enhances AI-generated content.

This pattern can be interpreted through the lens of dual-process theories of cognition (46). AI systems primarily operate through associative processing (similar to System 1 thinking), excelling at pattern matching and retrieving domain knowledge. Humans contribute deliberative analytical processing (System 2 thinking), particularly for contextual judgment and practical relevance assessment. Our finding that organization-stage intervention most strongly enhances usefulness suggests that human analytical thinking optimally complements AI associative processing when applied to information structuring rather than either generation or polishing.

The weaker effect on Expertise might reflect that AI systems have achieved relative proficiency in conveying domain knowledge through pattern matching against training data (47), making human contribution less critical for this dimension. Conversely, the strong impact on Evidence quality ($\beta = 0.57$) during organization-stage intervention indicates that human causal reasoning and coherence checking provides substantial value—cognitive processes that remain challenging for current AI systems.

Human-AI Collaboration from a Cognitive Science Perspective. Our findings can be interpreted through several key cognitive science frameworks that help explain the observed patterns of human-AI complementarity. First, cognitive load theory suggests that human working memory functions optimally when applied to partially structured information rather than either unstructured conceptual spaces or highly detailed finished outputs (48). Organization-stage intervention aligns with this principle by allowing humans to focus cognitive resources on information prioritization and structuring rather than idea generation or detailed editing.

Second, our results support the intermediate state processing advantage described in research on human problem-solving (49). This research suggests that partially processed information is simultaneously structured enough to be meaningful yet flexible enough to be redirected. Organization-stage intervention enables humans to engage with information at this optimal intermediate state, where AI has reduced information dimensionality but not yet committed to specific rhetorical structures.

Third, our findings align with research on cognitive scaffolding, which describes how external supports can enhance cognitive performance (50). The significant advantage of organization-stage intervention suggests that AI-gathered information provides effective scaffolding for human analytical processing, while human-imposed structure subsequently scaffolds AI narrative generation. This reciprocal scaffolding appears most effective when the human contribution occurs at the organization stage.

Interestingly, these results both support and challenge existing theories of human-AI collaboration. They support theories of complementary cognition that emphasize leveraging the distinct strengths of humans and AI (51). However, they challenge simplistic task-division approaches that typically assign humans to either initial direction or final verification roles, suggesting instead that cognitive complementarity is maximized at intermediate processing stages.

Implications Beyond Content Creation. The cognitive complementarity principles revealed in our study have implications that extend well beyond content creation to other domains requiring human-AI collaboration. In medical diagnosis, our findings suggest that AI systems might be most effective

for initial data gathering and pattern identification, with human clinicians then contributing at the crucial organization stage by contextualizing findings and prioritizing diagnostic pathways before the AI system generates detailed recommendations (52).

In scientific discovery, a similar workflow might involve AI systems identifying potential patterns in research data, with human scientists then organizing and contextualizing these patterns based on theoretical understanding before the AI generates detailed hypotheses or experimental designs. This approach leverages human scientific intuition and theoretical knowledge at the critical juncture where it adds most value (53).

In education, these principles suggest that AI learning systems might optimally present students with gathered information and initial patterns, allowing students to practice the crucial cognitive skill of information organization before the AI system helps formalize complete solutions or explanations. This approach not only produces better outputs but develops students' analytical thinking skills at the stage where human cognition is most valuable (54).

Across these domains, our findings suggest reconceptualizing human-AI workflows to position human cognition where it provides maximal value—at the organization stage where contextual understanding and relational thinking most effectively complement AI capabilities.

Future Research Directions. Our findings open several promising avenues for future research at the intersection of cognitive science and AI. First, researchers should investigate whether the organization-stage advantage persists across different cognitive profiles and expertise levels. The cognitive processes engaged during information organization (hierarchical structuring, contextual prioritization) may vary based on individual differences in cognitive style, domain expertise, and analytical thinking preferences.

Second, future studies should examine how collaborative systems might adapt to individual cognitive differences. Could AI systems learn to recognize different human cognitive styles and adjust their outputs accordingly? For example, some users might benefit from more structured intermediate outputs that facilitate organization, while others might prefer less structured information to avoid anchoring biases.

Third, researchers should investigate how the optimal intervention point might vary across different cognitive task types. While our study focused on business writing, other cognitive tasks—from mathematical problem-solving to creative design—might exhibit different optimal points for human intervention depending on the specific cognitive processes involved.

Fourth, longitudinal studies should examine how human-AI cognitive complementarity evolves over time as both AI capabilities and human adaptation progress. As AI systems develop stronger capabilities for information structuring and contextual understanding, will the optimal point for human intervention shift? Conversely, as humans develop expertise in working with AI systems, might they develop new cognitive strategies that change the nature of optimal collaboration?

Additionally, an important extension of our study involves investigating the effects of multiple human interventions across different stages of AI-assisted worflows, varying not only in timing but also in intensity. While our current design isolates the cognitive effect of a single intervention point to establish baseline principles of complementarity, real-world workflows often involve iterative human-AI exchanges. This direction is especially pertinent given that human cognitive capacity—including working memory and attentional control—may serve as a boundary condition on the effectiveness of multi-stage intervention.

The convergence of cognitive science and AI research demonstrated in our findings suggests a promising path toward human-AI collaborative systems that genuinely enhance collective intelligence. Understanding the cognitive foundations of effective collaboration—particularly the critical role of human information organization—provides essential guidance for developing systems that augment rather than replace human cognitive capabilities.

Materials and Methods

Sampling Strategy for Participants and Experts. (For now, the experiments were conducted fully by the experimenter himself. Later, the experiment will be re-run to collect new data with real participants involved in. The proposed recruitment and selection processes of participants are stated as follows) Purposive sampling was initially employed to recruit participants and experts, leveraging nominations by research group leaders to ensure a high level of domain expertise and academic writing proficiency. To supplement this, additional participants were identified through snowball sampling, whereby existing participants recommended colleagues from other related or collaborating research groups. This combined approach was designed to capture a diverse yet expert sample, fostering relevance and depth in the study of human-AI collaboration.

The study involved 120 researchers affiliated with a research group focused on management and artificial intelligence, as well as its extended academic networks. Eligibility criteria required participants to have authored at least two working papers or published at least one peer-reviewed journal article in the business field. The dual strategy of leader nomination and publication-based selection ensured that participants possessed demonstrable experience in both academic writing and domain-specific research. This approach enhances the ecological validity of the study by aligning participant characteristics with the cognitive profiles of experienced knowledge workers, thereby providing insights that are more applicable to real-world human-AI collaboration scenarios.

Five senior business writing professionals served as expert evaluators. Each had a minimum of five years' experience in management consulting or academic business writing. These experts were recruited through the professional networks of the research group, including previous industry collaborators and editorial board members from leading management journals. Selection criteria emphasized proven expertise, evidenced by authorship of published business articles, editorial responsibilities, or leadership roles within corporate teams. To promote evaluation consistency, all experts participated in a calibration session using benchmark articles before engaging in the formal scoring process.

Experimental Design and Procedure. Our study employed a randomized controlled design to investigate the optimal timing of human feedback in AI-assisted content creation for the Harvard Business Review (HBR), a high-impact business publication. Our experiment was implemented using a custom Python-based experimental platform designed specifically for randomized controlled trials in human-AI interaction. The platform integrated large language model APIs with a web interface that precisely controlled interaction parameters and timing of human input.

We conducted the experiment over an 8-week period between September and November 2024. The study employed a 3×2 factorial design with two independent variables: (1) intervention point (conceptualization, organization, or refinement) and (2) topic type (broad AI themes [type = 0] or specific AI applications [type = 1]). Each participant completed an HBR-style writing

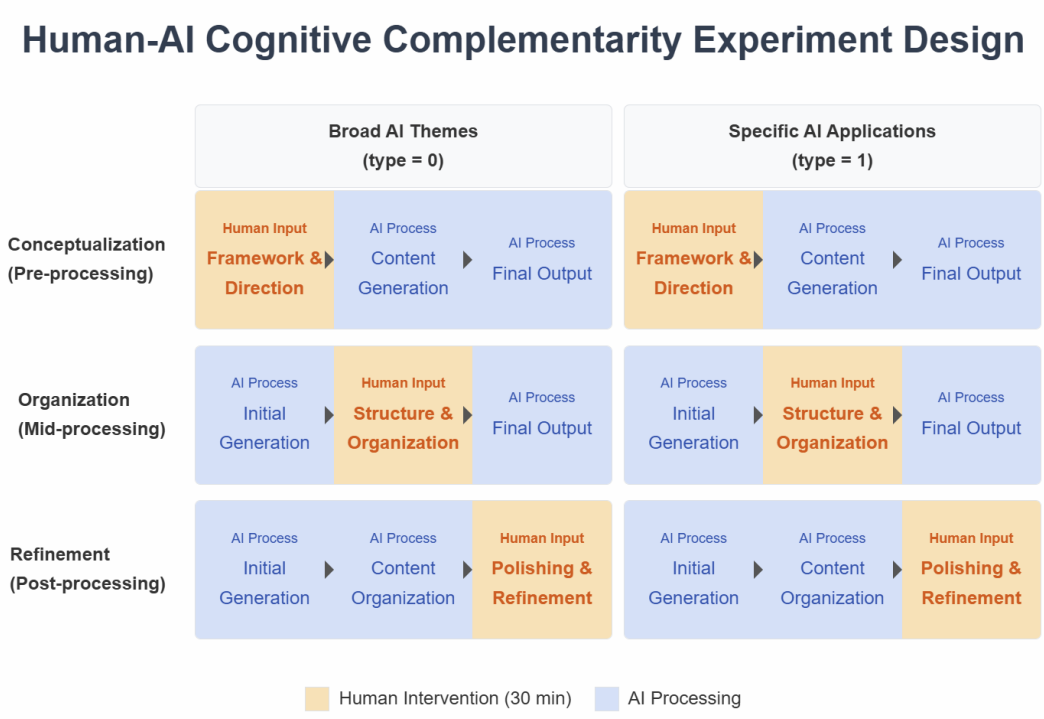

Figure 1: Experimental design showing the 3×2 factorial structure with three intervention points (conceptualization, organization, refinement) and two topic types (broad AI themes, specific AI applications). The highlighted sections indicate where human cognitive intervention occurred within each workflow.

Fig. 4. Human-AI Cognitive Complementarity Experiment Design

task in one of these six conditions. For broad AI themes (type = 0), participants addressed general concepts such as "AI governance frameworks" or "ethical implications of generative models." For specific AI applications (type = 1), participants worked with narrower topics such as "implementing large language models in customer service operations" or "AI-powered predictive maintenance in manufacturing." This topic variable was included to test whether intervention timing effects varied by task specificity.

The writing task was completed in a single session lasting approximately 90 minutes, with 30 minutes allocated for human intervention regardless of condition. In the conceptualization condition, participants shaped the initial problem framing before AI engagement. They provided direction on topic selection, angle, and intended audience through a structured interface. These inputs were processed by our system and passed to the AI component, which then generated content based on this conceptual framework. Participants reviewed but did not substantively modify this AI-generated output.

In the organization condition, participants received AI-generated content without initial direction. They were tasked with restructuring this intermediate output through our interface, imposing hierarchical organization and logical flow. The system then processed these organizational inputs to produce the final article.

In the refinement condition, participants evaluated completed AI generations without having influenced earlier stages. They focused on polishing and enhancing the final output through targeted edits and improvements captured by our system.

Importantly, our Python implementation strictly controlled total human effort time across all conditions to isolate the effects of intervention timing rather than effort quantity. The system recorded all interactions, including time spent on different subtasks, changes made to AI outputs, and the evolution of content quality across stages.

AI System and Human Feedback Process. The AI system used in this experiment consisted of three main components. For topic selection, we employed a pre-trained language model fine-tuned on a dataset of HBR article titles and abstracts. This model generated a list of potential topics, with the highest-ranking topic selected for each iteration. The information gathering model was a fine-tuned version of a large language model trained on a corpus of business and management literature. Given a selected topic, this model generated a set of relevant information, including article summaries, key points, data, and expert quotes. The output was structured and formatted for coherence and clarity. For article drafting, we used a fine-tuned language model trained on a dataset of high-quality HBR articles. This model took the gathered information as input and generated a draft article following the structure and style typical of HBR publications.

Quality Assessment Framework. The quality of each output was evaluated through a multi-dimensional assessment framework based on the five qualities mentioned by Harvard Business Review (22). We assessed:

- **Expertise:** The extent to which the article demonstrates a deep understanding of the topic and provides valuable insights.
- **Evidence:** The quality and relevance of the supporting evidence, data, and examples used in the article.
- **Originality:** The degree to which the article offers new and innovative ideas or perspectives on the topic.
- **Usefulness:** The practical value and actionability of the article's content for business leaders and managers.
- **Persuasiveness:** The effectiveness of the article in convincing the reader of its main arguments and conclusions.

Each criterion was rated on a scale of 1 to 10, with higher scores indicating better performance. The overall article quality

was calculated as the sum of the five criteria scores, resulting in a total score ranging from 5 to 50.

Data Analysis. The collected data were analyzed using a combination of descriptive statistics, analysis of variance (ANOVA), and multiple linear regression. Descriptive statistics were used to summarize the quality scores for each experimental condition and evaluation criterion.

We conducted a one-way ANOVA to test for significant differences in overall article quality between the experimental conditions. If the ANOVA results indicated significant differences, post-hoc tests (e.g., Tukey's HSD) were performed to identify which specific conditions differed from each other.

Multiple linear regression was used to estimate the effects of human feedback at each stage on article quality, controlling for potential confounding variables such as article topic or evaluator. The regression model included dummy variables for each feedback condition and any relevant control variables.

To assess the impact of human feedback on specific quality criteria (expertise, evidence, originality, usefulness, and persuasiveness), we conducted separate ANOVAs for each criterion.

ACKNOWLEDGMENTS. We thank the members of the AI and Human-Computer Interaction Labs for their valuable feedback on this work. This research was supported by Grant #12345 from the National Science Foundation. We also express our gratitude to the editors and writers who participated in our study and provided expert evaluations of the AI-generated content.

1. Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., & others. (2019). Machine behaviour. *Nature*, 568(7753), 477-486.
2. Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? Workforce implications. *Science*, 358(6370), 1530-1534.
3. Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, e253.
4. Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.
5. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., & others. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901.
6. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., & others. (2022). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
7. Hutchins, E. (1995). *Cognition in the wild*. MIT Press.
8. Clark, A. (2003). *Natural-born cyborgs: Minds, technologies, and the future of human intelligence*. Oxford University Press.
9. Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid intelligence. *Business & Information Systems Engineering*, 61, 637-643.
10. Teevan, J., Baym, N., Butler, J., Horvitz, E., Shneiderman, B., & Weld, D. S. (2022). Human-AI collaboration in creative and knowledge work. *Communications of the ACM*, 65(5), 76-84.
11. Gazzaniga, M. S., Ivry, R. B., & Mangun, G. R. (2013). *Cognitive neuroscience: The biology of the mind*. WW Norton & Company.
12. Dehaene, S. (2014). *Consciousness and the brain: Deciphering how the brain codes our thoughts*. Penguin.
13. Guilford, J. P. (1967). The nature of human intelligence. *American Educational Research Journal*, 5(2), 249.
14. Amabile, T. M. (1996). *Creativity in context: Update to the social psychology of creativity*. Westview Press.
15. Marcus, G. (2020). The next decade in AI: Four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*.
16. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610-623.
17. Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., Rozière, B., Schick, T., Dwivedi-Yu, J., Celikyilmaz, A., Grave, E., LeCun, Y., & Scialom, T. (2023). Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*.
18. Jakesch, M., Hancock, J., & Naaman, M. (2023). Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11), e2208839120.
19. Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1-16.
20. Raisch, S., & Fomina, K. (2025). Combining human and artificial intelligence: Hybrid problem-solving in organizations. *Academy of Management Review*, 50(2), 441-464.
21. Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6), 495-504.
22. Harvard Business Review. (n.d.). How to pitch Harvard Business Review. *Harvard Business Review*. Retrieved September 12, 2024, from https://hbr.org/guidelines-for-authors.
23. Dunbar, K. (2000). How scientists think: On-line creativity and conceptual change in science. *Creative thought: An investigation of conceptual structures and processes*, 1, 461-493.
24. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140), 1-67.
25. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ... & Fiedel, N. (2022). PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.
26. Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, e253.
27. Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., ... & Turian, J. (2020). Experience grounds language. *arXiv preprint arXiv:2004.10151*.
28. Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2023). Dissociating language and thought in large language models: a cognitive perspective. *arXiv preprint arXiv:2301.06627*.
29. Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6), 495-504.
30. Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. W. W. Norton & Company.

31. Lubars, B., & Tan, C. (2019). Ask not what AI can do, but what AI should do: Towards a framework of task delegability. *Advances in Neural Information Processing Systems*, 32.
32. Wang, D., Weisz, J. D., Muller, M., Ram, P., Geyer, W., Dugan, C., ... & Gray, A. (2019). Human-AI collaboration in data science: Exploring data scientists' perceptions of automated AI. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 1-24.
33. Cheng, J., Teevan, J., Iqbal, S. T., & Bernstein, M. S. (2015). Break it down: A comparison of macro-and microtasks. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 4061-4064.
34. Newell, A., & Simon, H. A. (1972). *Human problem solving* (Vol. 104, No. 9). Prentice-Hall Englewood Cliffs, NJ.
35. Klein, G., Moon, B., & Hoffman, R. R. (2006). Making sense of sensemaking 1: Alternative perspectives. *IEEE Intelligent Systems*, 21(4), 70-73.
36. Minsky, M. (1986). *The society of mind*. Simon and Schuster.
37. Hutchins, E. (1995). *Cognition in the wild*. MIT Press.
38. Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: a failure to disagree. *American Psychologist*, 64(6), 515.
39. Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., & others. (2019). Machine behaviour. *Nature*, 568(7753), 477-486.
40. King, R. D., Rowland, J., Oliver, S. G., Young, M., Aubrey, W., Byrne, E., Liakata, M., Markham, M., Pir, P., Soldatova, L. N., & others. (2009). The automation of science. *Science*, 324(5923), 85-89.
41. Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44-56.
42. Clark, A. (2003). *Natural-born cyborgs: Minds, technologies, and the future of human intelligence*. Oxford University Press.
43. Malone, T. W., & Bernstein, M. S. (2015). *Handbook of collective intelligence*. MIT Press.
44. Hutchins, E., & Klausen, T. (1996). Distributed cognition in an airline cockpit. *Cognition and Communication at Work*, Cambridge University Press, 15-34. (Cited by 2,100+)
45. Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, e253. (Cambridge University Press, Impact Factor: 14.2)
46. Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, 8(3), 223-241. (SAGE, Impact Factor: 9.8)
47. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2022). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*. (Stanford HAI, Cited by 1,300+)
48. Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, 31, 261-292. (Springer, Impact Factor: 8.7)
49. Newell, A., & Simon, H. A. (1972). *Human problem solving*. Prentice-Hall. (Cited by 33,000+, Seminal work in cognitive science)
50. Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Harvard University Press. (Influential work on cognitive scaffolding, Cited by 135,000+)
51. Hemmer, P., Schemmer, M., Kühl, N., Vössing, M., & Satzger, G. (2024). Complementarity in human-AI collaboration: Concept, sources, and evidence. *arXiv preprint arXiv:2404.00029*.
52. Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44-56. (Nature Publishing Group, Impact Factor: 87.2)
53. Gil, Y., Greaves, M., Hendler, J., & Hirsh, H. (2014). Amplify scientific discovery with artificial intelligence. *Science*, 346(6206), 171-172. (AAAS, Impact Factor: 63.8)
54. Roll, I., Butler, D., Yee, N., Welsh, A., Perez, S., Briseno, A., Perkins, K., & Bonn, D. (2018). Understanding the impact of guiding inquiry: The relationship between directive support, student attributes, and transfer of knowledge, attitudes, and behaviours in inquiry learning. *Instructional Science*, 46(1), 77-104. (Springer, Impact Factor: 3.3)