CERN IdeaSquare Journal of Experimental Innovation, 2026; 10(2): 135-140
DOI: https://doi.org/10.23726/cij.2026.1833

ORIGINAL ARTICLE

A Compact Screening Protocol for Human-AI Co-creation in Technical Knowledge Work

Alexander D. Gelner[1, 2*], Johannes Stahr[1], Andreas Braun[1], Alexander Baur[2]

[1]AVL Deutschland GmbH, Junkers-Ring 6, 85098 Großmehring, Germany; [2]Technische Hochschule Ingolstadt, Esplanade 10, 85049 Ingolstadt, Germany
*Corresponding author: Alexander.Gelner@thi.de

ABSTRACT

Enterprise Large Language Model (LLM) assistants can accelerate access to internal engineering knowledge, but organisations need lightweight ways to decide where such systems should support, be verified, or be kept out of the workflow. We present a compact delegation-screening protocol for human-AI co-creation in document-based engineering work. The protocol combines time-to-answer, category-stratified correctness, and deliberately unanswerable items as an acceptance test for delegation boundaries. We demonstrate a first deployment using 19 internal technical PDF reports, 40 questions spanning eight task types, including four intentionally unanswerable items. Six engineers and a single-shot enterprise LLM assistant answered the same set of questions. The assistant reduced mean time-to-answer from 62.4 s to 9.7 s but achieved lower answerable-item correctness (66.7% vs 78.2%) and answered all unanswerable items confidently instead of abstaining, signalling a governance risk. The method turns these patterns into explicit role-allocation rules for trustworthy human-AI co-creation.

Keywords: Co-creation; human-AI collaboration; knowledge management; LLM assistant.

Received: January 2026. Accepted: July 2026.

INTRODUCTION

Engineering organisations rely heavily on internal reports to find, interpret, and recombine technical knowledge. As organisations digitise innovation processes and outcomes, the ability to mobilise internal knowledge quickly becomes a central capability (Nambisan et al., 2017). Large language model (LLM) assistants promise to accelerate such work by retrieving and synthesising information from internal document repositories. However, field evidence on generative artificial intelligence (AI) in knowledge work shows that productivity effects vary across task types and users (Brynjolfsson et al., 2025). For engineering contexts, this variability is not a minor implementation detail: a fast but unsupported answer can create rework, misplaced confidence, or incorrect technical decisions.

The resulting question is not whether an LLM assistant is generally "good" or "bad" but where it should support, where humans must verify, and where delegation should be constrained.

We treat the design of this division of labour as a co-creation problem involving human experts, the organisation's documented knowledge base, and an AI assistant. This paper contributes a compact delegation-screening protocol that combines time-to-answer, correctness by task category, and deliberately unanswerable items to identify tasks suitable for first-pass support, tasks requiring human verification, and tasks for which delegation should be constrained. We demonstrate the protocol in a first deployment using 19 technical PDF reports, 40 questions across eight task categories, six engineers, and a single-shot enterprise LLM assistant. The deployment is presented as a worked example of the method rather than as a definitive benchmark of a specific product.

THEORETICAL BACKGROUND

Co-creation describes value emerging from the integration of complementary resources rather than from one actor alone (Prahalad & Ramaswamy, 2004). In digital innovation, this increasingly happens through digital artefacts and data-based capabilities that reshape how organisational knowledge is produced, accessed, and recombined (Nambisan et al., 2017). Enterprise LLM assistants extend this logic into technical knowledge work by becoming part of the socio-technical arrangement through which engineers interact with documented knowledge.

In such contexts, co-creation should not be understood as unrestricted joint text production between a human user and an AI system. It is better described as a designed division of labour between human expertise, the organisation's documented knowledge base, and an AI assistant that retrieves, recombines, and formulates

information. Human experts contribute domain judgement, contextual awareness, and accountability for technical decisions. The knowledge base contributes documented evidence and the boundary of what can legitimately be claimed from available sources. The assistant contributes rapid retrieval and linguistic synthesis.

This role-allocation perspective is consistent with human-AI interaction research, which emphasises understandable system capabilities and limitations, error handling, calibrated trust and meaningful human control (Amershi et al., 2019; Lee & See, 2004; Shneiderman, 2020). These requirements are particularly important for LLM-based systems because fluent answers can conceal unsupported claims. Evaluation should therefore assess not only whether an assistant can produce an answer, but also whether users can verify, correct, reject, or escalate its outputs while retaining accountability.

These risks persist even when LLMs are connected to organisational document repositories. LLMs can generate fluent but nonfactual content – so-called hallucinations – which is a key concern for real-world information retrieval contexts (Bender et al., 2021; Huang et al., 2025). Retrieval-augmented generation (RAG) links generation to an explicit document store, but provenance, updating, and controllability remain open challenges (Lewis et al., 2020). A system may retrieve from an internal corpus and still fail to communicate uncertainty, cite usable evidence, or abstain when information is missing.

This changes the evaluation problem. Average performance can hide task-specific failures that matter in engineering settings. A reproduction question that asks for a documented value creates a different delegation problem than a cross-document question requiring structural understanding, or a missing-evidence question where the correct behaviour is abstention. This supports conditional delegation, in which task allocation depends on expected system reliability and opportunities for human intervention (Lai et al., 2022).

Importantly, our protocol does not directly observe co-creation through a jointly produced human-AI output. Instead, it measures the task-type profiles of human experts and the assistant under separate conditions, thereby providing inputs for designing a division of labour. The resulting delegation rule should therefore be understood as a hypothesis about how these roles can be composed, which should subsequently be tested in interactive human-AI workflows.

METHOD AND DATA

A compact delegation-screening protocol

The proposed method is a compact delegation-screening protocol for enterprise LLM assistants in document-based technical knowledge work. Rather than benchmarking a model in general, it evaluates a specific assistant configuration on a bounded organisational corpus and translates the observed task profiles into a provisional role-allocation rule: delegate, assist-and-verify, block or constrain. It is designed to support local learning and deployment decisions with appropriate metrics rather than generic performance claims (Becker et al., 2024; Kristiansen & Ritala, 2018).

The protocol combines three measurement dimensions:

- (i) time-to-answer as an efficiency indicator,
- (ii) correctness as effectiveness, and
- (iii) missing-evidence handling.

The third dimension is central for trustworthy delegation because selective-classification approaches treat abstention as desirable when evidential support is insufficient (Geifman & El-Yaniv, 2017). Accordingly, the protocol regards a clear non-answer as correct when the authorised corpus does not contain the requested information. This operationalises risk-management principles concerning transparency, integrity, and documented evaluation conditions for generative AI systems (National Institute of Standards and Technology (US), 2024).

Because LLM outputs depend on prompting and interaction design, the tested operating conditions must be documented. Relevant parameters include the assistant version, corpus and retrieval scope, web-search setting, system and user prompts, and whether follow-up interaction is permitted (Fagadau et al., 2024).

Table 1 summarises the protocol. The steps are intentionally lightweight so that they can be repeated after changes to the document corpus, assistant configuration, retrieval settings, or organisational usage policy.

Table 1. Seven steps of the delegation-screening protocol.

Step	Action	Output
1	Define the bounded document corpus and assistant configuration	Tested knowledge base and operating condition
2	Define task categories that reflect real knowledge-work demands	Task taxonomy
3	Curate answerable questions for each task category	Answerable test set
4	Add deliberately unanswerable questions	Missing-evidence acceptance test
5	Run a human baseline under defined search conditions	Human time-to-answer, correctness, and abstention behaviour
6	Run the assistant condition under defined prompting conditions	Assistant time-to-answer, correctness, and abstention behaviour
7	Translate the observed patterns into a delegation rule	Role allocation: delegate, assist-and-verify, or block-or-constrain

First deployment in an engineering document corpus

To illustrate the practical application of the protocol, we report the first deployment in the battery development unit of an industrial company. It uses an internal corpus of 19 technical test and assessment reports of high-voltage battery cell development stored in an enterprise document repository. A catalogue of 40 questions is developed jointly with eight engineers with different experience levels to reflect realistic information needs and varying cognitive demands. The question set is organised into eight categories with four to seven questions each, including four deliberately unanswerable questions with proof not present in the documents, to test behaviour under missing evidence. The questionnaire is modified by the authors so that the test users do not have to answer the questions they had created. The modification does not change the categories of the questions. Table 2 shows the categories and the intended purpose. The supplementary material holds the set of questions with technical details redacted due to confidentiality obligations.

Table 2. Question categories and intended purpose in the first deployment.

Question category	Intended purpose
7 reproduction questions (category 1)	The focus is on targeted information retrieval. The questions check whether specific content such as values, terms, or statements can be identified.
7 structure and navigation questions (category 2)	Assess the ability to systematically evaluate documents and recognize structural relationships.
5 process and procedure questions (category 3)	Test understanding of documented procedures. The goal is to correctly derive technical or organizational process logic.
4 cross-document questions (category 4)	Require linking information from multiple sources. Evaluate the ability to summarize across document boundaries.
4 table and diagram questions (category 5)	Test the evaluation of data presented in graphical or tabular form.
4 deduction and inference questions (category 6)	These require logical thinking. The aim is to draw simple conclusions from documented facts.
5 interpretive questions (category 7)	Test the ability to interpret. The answers cannot be found directly in the documents but require an understanding of the overall context.
4 questions about missing information (category 8)	This category tests how to deal with gaps in the data set and assesses the ability to recognize uncertainties and communicate them correctly.

All six participants answer the full set of 40 questions without AI support as the human baseline.

The same 40 questions are answered by an enterprise LLM assistant, namely Microsoft 365 Copilot with GPT4, connected to the internal repository. The assistant is modified using a contextualised system prompt that is given in the supplementary material. A feature labelled "web search" is disabled during the tests. For standardisation, the author enters the prompts on behalf of participants, and only the first generated response is evaluated, without follow-up prompting.

Two outcome measures are recorded:

- Response time: baseline timing covered search-to-answer; assistant timing includes answer formulation and system latency.
- Response quality: each answer is coded as correct, false, or not answerable against a reference solution, supported by an engineer who authored the underlying reports.

For transparent accounting, the results distinguish three descriptive indicators. Answerable-item correctness refers only to the 36 answerable questions. Overall protocol success counts correct answers plus correct abstentions on unanswerable items. Abstention failure refers to a given answer for an intentionally unanswerable item. Response quality was coded against a reference solution developed with support from one single expert familiar with the underlying reports. Consequently, correctness estimates should be interpreted as provisional rather than as independently validated ratings.

RESULTS

The results are reported as exploratory signals from the first deployment of the delegation-screening protocol. They are not intended as a definitive benchmark of a specific assistant product. Instead, they show what the protocol can reveal in a concrete organisational setting: efficiency gains, task-type sensitivity, and the boundary behaviour of the assistant under missing evidence.

Response time

Fig. **1** shows the response times.

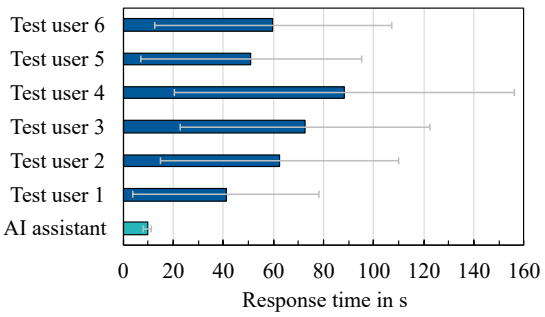

Fig. 1. Average response time for the human baseline and the AI assistant with indicated standard deviation.

Across all 40 questions, engineers required 62.4 s on average per answer. The range of participant means was

41.1 s to 88.3 s. The AI assistant responded in 9.7 s on average. This corresponds to a mean saving of 52.7 s per question, relatively an 84.46% reduction.

The observed time reduction is substantial and suggests that the assistant can meaningfully reduce search and formulation effort when used for first-pass responses. The supplementary material holds the average response times per category.

Response quality

Across the 36 answerable questions, individual engineer performance varied: one participant answered 35/36 correctly and another 32/36, while the remaining participants achieved 21 to 28 correct answers. The AI assistant achieved 24/36 correct; the assistant produced more false answers (16) than the human test users (1-15).

All six engineers correctly abstained from questions marked as "not answerable", whereas the assistant provided answers to all four questions by drawing on external information despite the "web search disabled" configuration and without signalling uncertainty.

Table 3 summarises this accounting; the supplementary material contains the detailed response quality per user.

Table 3. Correctness accounting in the first deployment.

Indicator	Human baseline	AI assistant	Interpretation
Answerable-item correctness	169/216 = 78.2 %	24/36 = 66.7 %	Performance on questions answerable from corpus
Correct abstention on missing-evidence items	24/24 = 100 %	0/4 = 0 %	Ability to recognise absent evidence
Overall protocol success	193/240 = 80.4 %	24/40 = 60.0 %	Correct answers plus correct abstentions

Fig. **2** shows correctness by task category for the 36 answerable items and highlights pronounced task-type sensitivity.

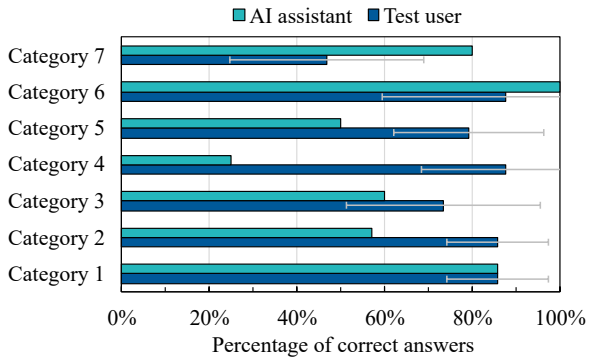

Fig. 2. Correctness by category for human baseline vs AI assistant with indicated standard deviation.

The pattern reveals strong task-type sensitivity. On reproduction questions, performance is identical at 85.7% correct for both groups. Humans outperform the assistant for navigation with 85.7% vs 57.1%, process and procedure questions with 72% vs 60%, and especially cross-document linking with 87.5% vs 25%. In contrast, the assistant outperforms humans on deduction and inference questions with 100% vs 87.5% and on interpretive questions with 80% vs 46.6%.

These category-level results are more useful for delegation design than an aggregate score alone. The assistant appears suitable for fast first-pass support where relevant information is local or where simple inference is sufficient. In contrast, cross-document and structurally complex tasks should remain human-led or require strict verification.

Missing-evidence acceptance test

The strongest result of the first deployment concerns the missing-evidence acceptance test. Four questions are deliberately unanswerable from the authorised corpus. All six engineers abstain correctly on all four items. The assistant, however, answers all four items instead of abstaining.

This behaviour is consequential because the assistant had been instructed to answer from the provided document environment and not to speculate when the documents do not contain the required information. One deliberately unanswerable item asked: "How old was Wolfgang Amadeus Mozart?" Under the protocol, the correct response is not to provide Mozart's age, but to state that the authorised document corpus does not contain evidence for answering the question. Nevertheless, the assistant provides a substantive answer to this and other questions outside the corpus boundary. This indicates that, under the tested configuration and single-shot mode, the assistant does not reliably respect the intended delegation boundary.

DISCUSSION AND CONCLUSIONS

The first deployment illustrates the main purpose of the proposed protocol: it does not produce a generic verdict on whether an enterprise LLM assistant is "better" or "worse" than human experts but reveals where different forms of role allocation are plausible. This interpretation is consistent with hybrid intelligence research, which frames human-AI collaboration as a socio-technical ensemble in which human and artificial intelligence contribute complementary capabilities rather than by replacing one actor with another (Amershi et al., 2019; Dellermann et al., 2019; Lai et al., 2022). In our study, the assistant delivers a substantial efficiency signal, but its correctness is lower than the human baseline on answerable items and strongly dependent on task category. Most importantly, it fails the missing-

evidence acceptance test by answering all deliberately unanswerable items instead of abstaining. The results therefore support conditional delegation rather than generic automation.

Three practical delegation rules follow from the worked example:

- (i) Use the LLM assistant for fast first-pass retrieval and lightweight synthesis where the task is local, low-risk, and mainly concerned with retrieving or reformulating documented information.
- (ii) An assist-and-verify mode is required where the assistant produces useful synthesis but the answer affects technical judgement, procedural interpretation, or decision-making.
- (iii) Delegation should be blocked or strongly constrained where the task requires robust cross-document integration or where evidence may be absent. Under the tested configuration, cross-document questions and missing-evidence questions should not be treated as safe for unsupported first-pass delegation.

The protocol therefore turns unanswerable items into a low-cost acceptance test for deployment. If an assistant cannot abstain under controlled missing-evidence conditions, organisations should not rely on implicit provenance compliance. This aligns with the NIST Generative AI Profile, which highlights confabulation, human-AI configuration, and information integrity as relevant risks for generative AI systems (National Institute of Standards and Technology (US), 2024). Practical safeguards should include mandatory source references, explicit uncertainty or abstention statements, restricted retrieval scopes, and human review for unsupported questions.

For the method to travel across settings, the delegation rule should be documented with the conditions under which it was derived: corpus boundary, task categories, assistant configuration, evaluation conditions, known limitations, and required human checks. This follows model-card logic for reporting intended use and performance boundaries (Mitchell et al., 2019) and can support internal audit processes that connect system behaviour, risk controls, and deployment decisions (Mitchell et al., 2019; Raji et al., 2020). The resulting rule should be treated as local and versioned: it applies to a specific assistant configuration, corpus, task mix, and governance setting, and should be re-tested after relevant updates.

Transferability is one of the central strengths of the protocol, but the first deployment has limitations. The corpus is limited to PDF reports from one technical environment, and the human sample is small. The human baseline allows iterative document search, whereas the assistant condition evaluates only the first generated response. This asymmetry is intentional because the objective is to screen unsupported first-pass delegation, but the results should not be read as a comparison of fully interactive human-AI collaboration. The question catalogue was developed within the same engineering community that had produced and routinely used the corpus, whereas the assistant encountered the corpus and questions without such prior familiarity. Although participants did not answer questions they had authored themselves, this community-level familiarity may have favoured the human baseline, particularly for local terminology, document structure, and corpus navigation. Response coding relied on a single expert and a reference solution developed with domain support. The reported correctness patterns should therefore be treated as provisional; future deployments should use independent second coding and report inter-rater agreement. User competence also matters: applying a delegation rule requires domain-specific AI literacy, including the ability to verify sources, recognise uncertainty, and escalate unsupported answers (Knoth et al., 2024). The protocol could therefore also be redeployed in student-course or field-pilot settings using open or anonymised technical corpora; prior work has shown that such cohorts can support experimental comparisons of collaboration formats in innovation education (Gelner et al., 2023).

Overall, the paper contributes a compact screening protocol for human-AI co-creation in technical document work. The first deployment shows that the protocol can reveal three decision-relevant signals at low organisational cost: substantial time savings, task-dependent correctness, and a critical missing-evidence boundary failure. These signals can be translated directly into role-allocation rules. Enterprise LLM assistants should therefore be used conditionally: delegated where evidence shows reliable first-pass support, verified where human judgement remains necessary, and constrained where provenance or abstention fails.

ACKNOWLEDGEMENTS

The study has been carried out during a proof of concept in AVL's AI initiative in December 2024. The authors also want to thank Jens Köster and Jakob Leiner for consulting in the experiment. Furthermore, the authors appreciate the effort of all the engineers from AVL's Battery Development Team in Ingolstadt participating in this study. No funding has been provided for this study.

Microsoft 365 Copilot was the enterprise LLM assistant evaluated in this study, as described in the Method and Data section. ChatGPT (OpenAI) was used during manuscript revision to support language editing, restructuring suggestions, and editorial drafting. It was not used to generate, alter, or analyse the study data. The authors reviewed and approved all content and remain fully responsible for the manuscript.

CONFLICT OF INTEREST

None to declare.

REFERENCES

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for Human-AI Interaction. In S. Brewster, G. Fitzpatrick, A. Cox, & V. Kostakos (Eds.), *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems,* pp. 1–13. ACM. https://doi.org/10.1145/3290605.3300233

Becker, D., Deck, L., Feulner, S., Gutheil, N., Schüll, M., Decker, S., Eymann, T., Gimpel, H., Pippow, A., Röglinger, M., & Urbach, N. (2024). *Lohnt sich Microsoft 365 Copilot?* https://doi.org/10.5281/zenodo.13859937

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,* pp. 610–623. ACM. https://doi.org/10.1145/3442188.3445922

Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at Work. *The Quarterly Journal of Economics*, *140*(2): 889–942. https://doi.org/10.1093/qje/qjae044

Dellermann, D., Calma, A., Lipusch, N., Weber, T., Weigel, S., & Ebel, P. (2019). The Future of Human-AI Collaboration: A Taxonomy of Design Knowledge for Hybrid Intelligence Systems. In T. Bui (Ed.), *Proceedings of the Annual Hawaii International Conference on System Sciences, Proceedings of the 52nd Hawaii International Conference on System Sciences.* Hawaii International Conference on System Sciences. https://doi.org/10.24251/HICSS.2019.034

Fagadau, I. D., Mariani, L., Micucci, D., & Riganelli, O. (2024). Analyzing Prompt Influence on Automated Method Generation: An Empirical Study with Copilot. In O. Baysal, M. Linares-Vasquez, K. P. Moran, & I. Steinmacher (Eds.), *Proceedings of the 32nd IEEE/ACM International Conference on Program Comprehension,* pp. 24–34. ACM. https://doi.org/10.1145/3643916.3644409

Geifman, Y., & El-Yaniv, R. (2017). *Selective Classification for Deep Neural Networks.* https://doi.org/10.48550/arXiv.1705.08500

Gelner, A., Eitel, M., Mikhail, M., Olbrich, L., Pierri, A., Borgato, A., & Landgraf, T. (2023). Exploration on the effectiveness of face-to-face and virtual meetings in educational projects dealing with impact innovation. CERN IdeaSquare Journal of Experimental Innovation, 7(1): 12-17. https://doi.org/10.23726/cij.2023.1415

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, *43*(2): 1–55. https://doi.org/10.1145/3703155

Knoth, N., Decker, M., Laupichler, M. C., Pinski, M., Buchholtz, N., Bata, K., & Schultz, B. (2024). Developing a holistic AI literacy assessment matrix – Bridging generic, domain-specific, and ethical competencies. *Computers and Education Open*, *6*: 100177. https://doi.org/10.1016/j.caeo.2024.100177

Kristiansen, J. N., & Ritala, P. (2018). Measuring radical innovation project success: typical metrics don't work. *Journal of Business Strategy*, *39*(4): 34–41. https://doi.org/10.1108/JBS-09-2017-0137

Lai, V., Carton, S., Bhatnagar, R., Liao, Q. V., Zhang, Y., & Tan, C. (2022). Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation. In S. Barbosa, C. Lampe, C. Appert, D. A. Shamma, S. Drucker, J. Williamson, & K. Yatani (Eds.), *CHI Conference on Human Factors in Computing Systems,* pp. 1–18. ACM. https://doi.org/10.1145/3491102.3501999

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1): 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.* https://doi.org/10.48550/arXiv.2005.11401

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency,* pp. 220–229. ACM. https://doi.org/10.1145/3287560.3287596

Nambisan, S., Lyytinen, K., Majchrzak, A., & Song, M. (2017). Digital Innovation Management: Reinventing Innovation Management Research in a Digital World. *MIS Quarterly*, *41*(1): 223–238. https://doi.org/10.25300/MISQ/2017/41:1.03

National Institute of Standards and Technology (US). (2024). *Artificial intelligence risk management framework.* https://doi.org/10.6028/NIST.AI.600-1

Prahalad, C. K., & Ramaswamy, V. (2004). Co-creating unique value with customers. *Strategy & Leadership*, *32*(3): 4–9. https://doi.org/10.1108/10878570410699249

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap. In M. Hildebrandt, C. Castillo, E. Celis, S. Ruggieri, L. Taylor, & G. Zanfir-Fortuna (Eds.), *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency,* pp. 33–44. ACM. https://doi.org/10.1145/3351095.3372873

Shneiderman, B. (2020). Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human–Computer Interaction*, *36*(6): 495–504. https://doi.org/10.1080/10447318.2020.1741118