Crafting Future Scenarios with the Help of AI: Potentials of a Hybrid Delphi Expert Panel

Roland M. Mueller
Berlin School of Economics and Law, Germany
roland.mueller@hwr-berlin.de

Katja Thoring
Technical University of Munich, Germany
katja.thoring@tum.de

Hermann W. Klöckner
Anhalt University of Applied Sciences, Germany
hwkloeckner@icloud.com

Kai R. Larsen
University of Colorado Boulder, USA
kai.larsen@colorado.edu

Abstract

This paper examines the potential of ChatGPT to enhance the established Delphi method by providing additional AI-infused expertise. We investigated several aspects of the Delphi method independently: the integration of abstract AI-infused expertise perspectives, the generation of an AI "clone" (digital twin) of a human expert, the rating of scenarios through AI, and the capability of AI to iterate future scenarios and to provide qualitative feedback. The findings suggest that AI systems can augment a Delphi panel by providing new perspectives but cannot replace individual human experts and their respective expertise. The insights shall inform other researchers who want to conduct hybrid Delphi studies with AI-infused expertise. In that sense, with this paper, we aim to lay the foundation for a hybrid Delphi study method and suggest actionable recommendations.

Keywords: Futurology, Delphi Method, Scenario Planning, Artificial Intelligence, ChatGPT

1. Introduction

Addressing design for the future presents a ubiquitous challenge within the disciplines of Design and Information Systems (IS). Developing artifacts for future contexts can provide companies with a competitive advantage (Thoring et al., 2022) and might also help address the complex challenges the world is facing.

One important approach when designing for the future is working with scenarios. Future scenarios are typically developed and evaluated regarding their likelihood of becoming a reality through working with a panel of domain and futurology experts, for example, as part of a Delphi study (Linstone & Turoff, 1975; Thoring et al., 2022). A Delphi study assembles a panel of experts who are asked either to develop future scenarios or to rate existing scenarios presented to them. The goal of this approach is typically to reach a consensus among the experts regarding the scenarios. This can partly be achieved by quantitatively rating the scenarios or by actively improving them by providing qualitative feedback and suggestions for modifications.

However, such a Delphi study with experts is very time and cost-intensive since the experts have little time and are usually highly paid. Hence, this method is not very accessible and affordable for most researchers. This situation leads to the following research question that guided our study:

RQ: What is the potential of Generative AI (specifically ChatGPT-4 with its current capabilities) to augment a Delphi panel of human experts through an AI-infused approach?

This research question touches on various related aspects, such as: (1) Can we democratize access to collective expert knowledge through Generative AI? (2) Can we expand the established Delphi method by including "AI experts" that would bring in specific expertise and, hence, could serve as some sort of researcher triangulation? (3) Can we create a "digital twin" of a particular human expert for researchers to work with? Finally, (4) can ChatGPT provide comparable or even better qualitative feedback to improve the scenarios or inspire new ones? These questions will be explored and discussed in the remainder of this paper.

2. Theoretical Concepts

2.1. Future Studies

There are various established methods for future forecasting and scenario development: Trend

URI: https://hdl.handle.net/10125/107160
978-0-9981331-7-1

extrapolation and forecasting typically rely on existing data in the present that are projected into the future (Shell International BV, 2008). Such trend forecasting studies are frequently used in the IS discipline (Hovorka & Peter, 2019). The Delphi method is a systematic approach to developing future scenarios involving a panel of experts. It is based on the assumption that a collective opinion would outperform individual expertise (Linstone & Turoff, 1975). However, in the Information Systems (IS) discipline, systematic future studies are not well-established (Carmel et al., 2011). One possible reason for this situation could be the time and effort required to assemble a panel of experts.

An overview of futurology methods and approaches in relation to the IS field can be found in Hovorka and Peter (2019). Thoring, Mueller, et al. (2023) present a framework for developing, discussing, and evaluating future-oriented artifacts.

The goal of this paper is to explore the potentials of Generative AI, and ChatGPT in particular, to enhance a Delphi study by providing additional forms of expertise.

2.2. Human-AI Collaboration

While AI systems in the past were mainly used to automate routine tasks, nowadays, they "work jointly with humans like teammates or partners to solve problems" (Lai et al., 2021, p. 390). For example, in the healthcare context, a physician could be supported with diagnosis tasks by a clinical decision support system. According to Abedin et al. (2022, p. 691), the "social aspects of interactions between humans and AI systems remain under-researched."

Several authors investigate the potential of "AI as a teammate" in scientific, educational, and creative collaboration. Seeber et al. (2020) outline research opportunities that arise from collaboration with autonomous agents, whereas Ma et al. (2023) suggest that while AI has the potential to generate scientific content that is as accurate as human-written content, there is still a gap in terms of depth and overall quality. Elshan and colleagues suggest requirements for creative workshops in which teams are augmented with AI team members (Elshan et al., 2022; Siemon et al., 2022). Wilson and Daugherty (2018, p. 4) stress that AI has "the most significant impact when it augments human workers instead of replacing them." Among the identified aspects that could be enhanced through AI, they suggest decision-making processes, for example, by integrating digital twins.

2.3. Generative AI

Generative AI is a subdomain of Artificial Intelligence (AI). It can create a variety of outcomes (plain text, code, images, videos, 3D models, music, etc.) in response to so-called prompts. A prompt is typically a text-based input in which the user can provide instructions for the AI and descriptions of the desired result to be generated by the system. But prompts can also take other forms as input, for example, images, sound, or speech. The specifics of the prompt highly influence the quality of the outcome generated by the AI (Wang et al., 2023), which elicited the need for a new form of expertise: prompt engineering. The possibilities of Generative AI to augment human designers is discussed by Thoring et al. (2023).

ChatGPT is an AI chatbot based on GPT-4, enabling a variety of use cases, such as generating and interpreting texts or computer code, writing song lyrics, and composing music (OpenAI, 2023). A recent research paper authored by 43 experts from various disciplines attributes ChatGPT with the ability to generate sophisticated text indistinguishable from that produced by a human (Dwivedi et al., 2023). The authors acknowledge ChatGPT's potential to enhance productivity. This statement is in line with our own assumption that ChatGPT might be able to team up with and facilitate designers by providing expert-like input and feedback to enhance the crafting of future scenarios. For this research paper, we are particularly interested in ChatGPT, as it enables us to instruct the AI to evaluate and discuss text-based future scenarios.

2.4. Digital Twin

A digital twin is a digital representation of a particular physical product or system that can generate comparable information to its physical counterpart (Grieves, 2019). The concept originated in the engineering field, for example, to simulate, monitor, or test a specific machine's performance. The potential application of a digital twin ranges from simulating only small aspects of an entity to a user-controlled avatar with a memory and perhaps agential ability to collect data from a digital environment to a comprehensive "twin" of an entity.

A specific form of a digital twin is a so-called "human digital twin" (Bomström et al., 2022), which is a digital counterpart of a human being. This concept is particularly relevant in the healthcare field in the form of a digital twin of a patient (Bruynseels et al., 2018). One of the goals of this paper is to explore whether ChatGPT could serve as a digital twin of a human expert by mimicking his/her expertise and opinions. This might be achieved by prompting ChatGPT with specific information of the expert's characteristics and expertise descriptions. We want to explore the potential of this digital twin to act as a specific type of "AI-infused" expert.

2.5. Expertise

According to Ericsson (2003), reaching the expertise level of an "expert" requires over ten years of active engagement in a topic, including deliberate (i.e., reflected) practice. Based on this, experts to be included in the expert panel of a Delphi study should be at least ten years active in the respective domain for which their expertise is required.

In contrast to this notion of "expert knowledge," we refer to the concept of "collective intelligence" (Leimeister, 2010) that suggests that the combined "wisdom of crowds" (Surowiecki, 2004) outperforms individual expertise for certain kinds of problems.

We argue that Generative AI in general and ChatGPT specifically can act as an interface to such collective knowledge. The algorithm is within seconds able to access the published knowledge in a captured domain. This raises the question of whether the technology could augment (or even replace) human experts or provide us with additional verification and triangulation options. The question of how technologies will impact the work of human experts is discussed, for example, in Susskind and Susskind's book "The Future of the Professions" (2022).

With this paper, we aim to explore whether ChatGPT can provide us with access to a particular expertise that is encoded in a latent space of the embeddings of the published knowledge. A latent space is a compact mathematical representation that has been trained on a huge amount of explicit knowledge, ensuring that semantically similar points are located near each other (Arvanitidis et al., 2018). Appropriate prompts can create useful views on this collective knowledge base (Dwivedi et al., 2023).

3. Research Method

3.1 Research design

The investigation presented in this paper is based on existing data from a traditional (non-AI-supported) Delphi study with a human expert panel of 20 experts and two human moderators. The Delphi study involved initial scenario development, two rounds of scenario ratings and qualitative feedback, a resulting set of 23 iterated future scenarios, and a concluding workshop to discuss the results. Based on this data, we now explore potential AI support for some of the involved steps.

Figure 1 illustrates our research design. The right area shows the traditional Delphi study with four steps: (1) development of future scenarios, (2) expert rating of the scenarios (asynchronously), (3) qualitative feedback on the scenarios (asynchronously), (4) iteration of the scenarios, and (5) a concluding workshop. Steps 2 and 3 were performed in two iteration cycles and resulted in 23 iterated scenarios. The left area illustrates potential AI support for each step.

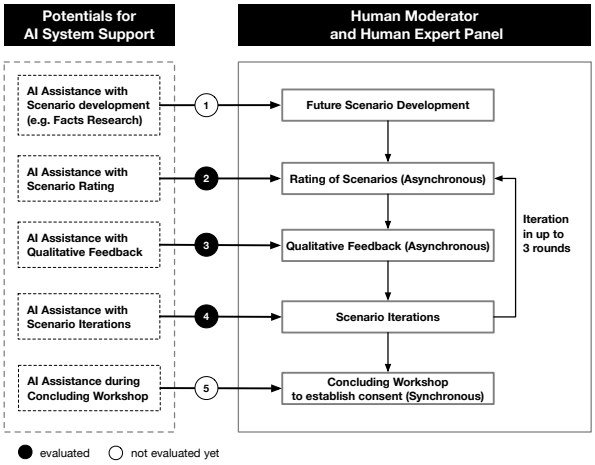

Fig. 1: Schematic overview of the research design.

For this study, we focused on steps 2, 3, and 4 only. The possibilities of the AI system to facilitate scenario development (step 1) and the concluding workshop (step 5) need to be investigated in future work. Potential AI support for steps 2, 3, and 4 are explored in two different ways:

First, we prompted ChatGPT-4 to rate the scenarios and provide feedback by taking the position of a specific (rather general) expertise from a specific discipline.

Secondly, we tried to create ChatGPT-4-based "digital twins" of individual experts. For this purpose, we prompted ChatGPT with the exact names of the experts from our previous Delphi study and instructed the AI system to consider any available published knowledge from that person to be found, for example, in their social media profiles. The reason for creating digital twins of our *existing* expert panel was to be able to compare the responses with each other. That way, we could infer whether the feedback from the AI system would be comparable to the one from the human experts. The more similar the answers are, the more certain we can be that an AI system could replace a human expert.

If we succeeded with our first goal (infusing general expertise into a Delphi study), this would allow future researchers to include additional expertise into a Delphi panel and to use responses from the AI system as some form of researcher triangulation.

If we succeeded with our second goal (creating digital twins of individual experts), researchers could include digital twins of particular experts – who are not

available for the human expert panel – into their Delphi study. This would be particularly relevant if very specific expertise is required (for example, by someone like Elon Musk), but that person is unavailable.

For the remainder of the paper, we use the term "AI expert" to address the more general AI-infused expertise (our first goal) and the term "AI digital twin" for the exact digital copy of an individual expert (our second goal). We do not claim that an AI-infused system resembles a real person with human properties. However, to illustrate the idea of creating a digital counterpart of general expertise or an individual expert, we use the term "AI expert" or "AI digital twin," respectively. In the following, we describe our explorative study design in more detail.

3.2 Baseline study

The Delphi study used as the baseline for our research was conducted in 2021. It addressed the "office of the future" and involved an expert panel of 20 domain and futurology experts from various disciplines. The expertise of the experts included futurologists and trend researchers, domain-specific experts like office planners, architects, and technology experts, as well as science fiction authors. They had an average of 19.3 years and a median of 15 years of professional experience. 23 future scenarios were developed mainly based on secondary research and imagination. Three thematic clusters were addressed: (1) The office after COVID-19, (2) the office for Generation Z, and (3) the office under the impact of emerging technologies. The envisioned time horizon was 10-15 years into the future. The scenarios were described by a title and a text of approximately 80 words.

For example, scenario 6 was titled "Visible Health" and suggested that employees' health would be discretely monitored and analyzed by the office infrastructure. Scenario 10 titled "Generation Clash," suggested that due to the aging population, increased retirement age, and the entry of Gen-Z into the workforce, various conflicts would emerge in the workplace. Scenario 20 was titled "Self-Driving Vehicles as the New Office" and suggests that self-driving cars could be used during the commute for administrative tasks and team meetings.

The developed scenarios were rated and discussed by the 20 experts and accordingly iterated further throughout two rounds of qualitative feedback and quantitative rating. Details about the entire project can be found in Thoring et al. (2022) and Thoring, Mueller, et al. (2023).

The resulting data (numeric ratings and qualitative feedback from real human experts) now allows us to test whether an AI-infused system could create comparable results and come up with useful feedback.

3.3 AI-infused inquiries

The second stage of the project was conducted in early 2023. We use the data from the Delphi study to compare the human experts' input with input from ChatGPT-4 regarding the developed scenarios. We prompted ChatGPT to evaluate the 23 existing scenarios and to take on the role of each of the respective experts. These "AI experts" have been created based on the expertise descriptions provided by the human experts in the original Delphi study.

Consequently, we were able to compare 20 human experts and 20 "AI experts" to each other. We compared how the AI-infused system rated the scenarios to their human counterparts, regarding scenario probability on a 5-point scale from -2 (very unlikely) to +2 (very likely). Qualitative feedback for each scenario was enquired by the AI system as well, and we asked to provide alternative versions of the scenario or suggest new ones, where appropriate (the same assignment was previously given to the human experts). In several rounds of iteration, we tried to improve the prompts addressed to ChatGPT in the sense that the provided responses would be more similar to those of the human experts. This would allow us to infer whether ChatGPT could become a valuable team member of a hybrid Delphi panel.

We conducted several rounds of analyses, for which we mainly used a combination of the ChatGPT-4 web interface and the ChatGPT-4 API in a Python programming environment. The latter allowed us to quickly conduct several inquiries (as outlined in the next section) with modified prompts. The results were compared against each other and the baseline from the human experts. The following areas of interest informed our research:

(1) The aggregate of all AI and all human expert ratings retrieved through a baseline prompt with abstracted expertise keywords and only limited context information
(2) Various modifications of the baseline prompt with more precise expertise keywords and adjustments of the ChatGPT temperature
(3) Pair-wise comparison of the human experts and their AI digital twin expert
(4) Comparison of the qualitative feedback provided by the human and the AI experts
(5) Qualitative analysis of scenario variations provided by the AI experts.

While human experts would access their individual expertise, AI experts would be prompted to access the Large Language Model (LLM) knowledge base.

3.4 Prompt Engineering

We started with an initial prompt (1) that only included abstract information regarding discipline and expertise (e.g., Architect, Office Planner, Science Fiction author, etc.) and only marginal information about the Delphi study. This initial prompt serves as a prompt template for further prompt engineering:

> *"You are an expert in [Expertise], who has been selected for a Delphi study. Please rate the following scenarios in terms of their likelihood of occurrence, from 'very unlikely' to 'rather unlikely' to 'neutral' to 'rather likely' to 'very likely.' Alternatively, you can indicate, "I cannot/will not evaluate this scenario." Please briefly explain your assessment of the scenarios below. Feel free to rephrase them if you wish [followed by the full text of the 23 scenarios]. Format the answer as a CSV, with the evaluation and justification in separate columns."*

For the subsequent steps, we (2) expanded the prompt by adding more details and instructions. For example, we included information about the topic of the Delphi study ("the office of the future") and the addressed time horizon (10-20 years from now). In the third step (3), we included the exact names of the experts, their positions, and their years of experience. We suggested accessing any available published knowledge to create a human digital twin of the experts. For each step, (4) randomly selected qualitative free-text responses were analyzed. Finally, we instructed ChatGPT to (5) provide three variations of selected scenarios that would be more interesting and speculative.

The ChatGPT API also allows the adjustment of the so-called 'temperature.' A lower degree (e.g., 0.3) would result in more obvious, conservative responses, while a higher degree (e.g., 0.9) would result in more speculative, surprising responses. We used a temperature of 0.6, 0.8, and 1.0, respectively.

3.5 Analysis

For the quantitative data analysis, we used a combination of spreadsheet visualization and evaluation and a Python Jupyter notebook with the following Python metrics libraries: Pingouin (Vallat, 2018) and Python Confusion Matrix (PyCM) (Haghighi et al., 2018).

For the qualitative data analysis of the free-text responses, two researchers read, discussed, and compared selected quotes. Their assessments were then aligned until they found an agreement. Additionally, the quotes from human and AI experts were compared quantitatively regarding their average length and how often they were provided. The results of the data analysis process are presented and discussed in the next sections.

4. Results

4.1. Results of the AI-infused panel

To evaluate whether the aggregate of the AI experts evaluated the scenarios in a similar way as the human experts, we calculated the average rating of the human experts per scenario. We compared them with the average rating of each scenario of the AI experts. The Pearson correlation coefficient of the average scenario ratings between AI experts and human experts is r=0.64 (n=23, p<0.001), which indicates moderate reliability between the group of AI experts compared to the group of human experts.

Figure 2 compares the average rating between human and AI experts per scenario. In 78% of the scenarios (18 of 23), human and AI experts had the same direction (sign) of the average rating. Figure 2 also shows that the average of the AI experts is, in most scenarios, more extreme on the positive and negative side. This is also shown by the average of the absolute average ratings, which is 0.45 for the humans and 0.93 for the AI. The paired t-test indicates a significant difference of the absolute average ratings ($t(22) = -5.34$; $p < 0.0001$).

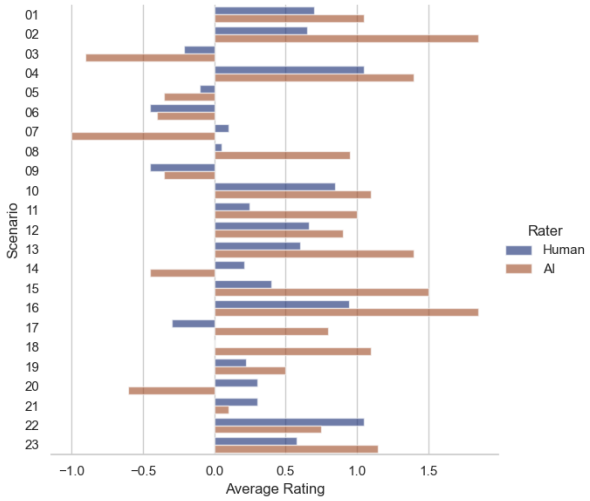

Fig. 2: Average Rating for Human and AI Experts with the Baseline Prompt

The standard deviation of the ratings of the AI experts was smaller than that of the human experts in all

scenarios. The average standard deviation per scenario for humans was 1.13, and for the AI experts, it was 0.47 ($t(22) = 11.281$; $p < 0.0001$). Therefore, AI experts tend to agree more with each other in each scenario, but the average ratings of the scenarios are more extreme than those from the human experts. Among all scenarios, the difference in the standard deviations between humans and AI is less severe: the standard deviation for the humans is 1.20, and for the AI experts it is 0.99.

Sablatzky (2022) observed that humans in a Delphi method are prone to a 'social desirability bias' and avoid extreme negative ratings and harsh feedback. However, our findings show that the AI experts are avoiding negative ratings far more often than the human experts. The overall distribution across all scenarios (see Figure 3) shows that AI experts tend to be rather positive and avoid the -2 (very unlikely) option. The AI experts only voted -2 (very unlikely) once out of the 420 votes, while the real experts voted -2 (very unlikely) 41 times. Overall, the AI experts voted more positively: the average rating is 0.58 (on a scale from -2 to +2), while human experts voted 0.32 on average. This indicates a more significant 'social desirability bias' of ChatGPT than of the human experts.

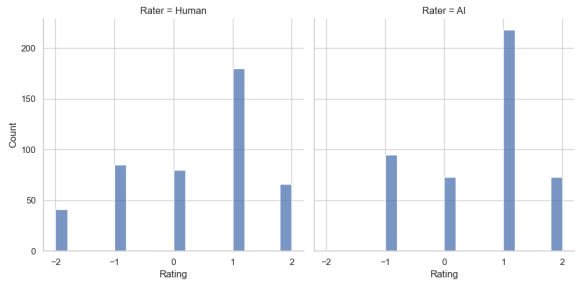

Fig. 3: Distribution of ratings for human and AI experts with the baseline prompt

Through prompt engineering (as outlined in Section 3.4), we tried to improve the average agreement between human and AI experts by adding more context information about the study and adjusting the temperature.

To assess whether AI experts' evaluation of the scenarios mirrored those of human experts, we computed the mean score for each scenario from the human experts. We compared these to the mean scores given by the AI experts for each scenario. A Pearson correlation coefficient was calculated to measure the association between the average scenario ratings from AI and human experts. This yielded a correlation of $r=0.45$ ($n=23$, $p<0.05$), still pointing towards a moderate correlation when comparing the AI expert group with the human expert group, but slightly worse than for the initial prompt.

4.3. Results of Digital Twin prompt

Finally, we compared the ratings of each individual human expert to their exact digital counterpart (the "AI Digital Twin", which was constructed as described in Section 3.4.)

To determine whether an AI digital twin expert gives a rating similar to that of its human counterpart, we matched each human expert's assessment with its AI digital twin (n=433 pairs). We considered the individual scale responses of the experts and their digital twins as separate ratings and computed the interrater reliability among them. With Cohen's kappa coefficient of $\kappa=0.12$ and Krippendorff's alpha coefficient of $\alpha=0.08$, both indices reflect low interrater concordance. When we treated the ratings as numerical values (ranging from -2 to +2) and computed a Pearson correlation coefficient, the resulting value was $r=0.11$ ($n=433$; $p<0.03$), indicating a poor level of agreement.

Therefore, we conclude that constructing a "human digital twin" based on the prompted name, expertise profile, and job details that successfully mimics the answers of the human twin is not yet possible.

4.4. Results of qualitative feedback

We systematically compared the qualitative feedback and explanations provided by the human and the AI experts based on selected quotes, which yielded the following insights:

The ChatGPT-based AI experts provided qualitative explanations in every instance. In contrast, the human experts only provided such additional feedback and explanations in roughly 25 percent of the cases. The qualitative explanations provided by the human experts tended to have a commentary nature and provide additional questions and hints. In contrast, the ones provided by the AI experts tended to justify the ratings and underpin them by providing context. The qualitative feedback provided by the human experts has also been noticeably shorter: On average, the human experts' response counts 12 words (often no replies at all), while the average word count of the AI experts was 35 words (never no reply).

Although typically shorter, the human experts' qualitative feedback appeared more concise. It provided additional reframing questions, as well as suggested sources, which the AI experts did not. For example, the human expert for 'Architecture Interior Architecture, Furniture, Office Planning, and Work Organization' commented on the scenario 'The office of the future is going to be neo-analog' with the following feedback: *"The latest available technology will always be used!"* The corresponding AI expert

replied, *"While Generation Z values climate neutrality and sustainability, it is unlikely they will completely reject digital media. The idea of a 'neo-analog office' is interesting but could bring challenges regarding efficiency and productivity."* This indicates the tendency of ChatGPT to be less questioning and opinionated than the human experts in their qualitative replies.

4.5. Results of scenario iterations

We randomly selected 10 original scenarios and instructed ChatGPT-4 to develop three variations for each, which equals a sample of 30 new scenarios developed by ChatGPT. Two researchers independently evaluated the 30 new scenarios based on the innovativeness and likelihood between +2 and -2. When we compare the direction of the rating as either negative (-2, -1) or positive/neutral (0, +1, +2), there was a Cohen's Kappa of 0.44 which indicates moderate agreement. The average innovativeness of the new scenarios developed by ChatGPT is 0.3. From the 30 variations, 10 scenarios were rated 'very innovative' (+2) by at least one of the annotators. In two cases, both annotators agreed with the 'very innovative' (+2) rating. In 8 variations, both annotators at least had a rating of 'innovative' (+1) or 'very innovative' (+2). This shows that ChatGPT is capable of instigating new ideas.

When we compare the likelihood of the 30 new scenarios based on the average rating of the two annotators and compare them with the average likelihood of the original scenarios rated by the 20 human experts, the likelihood was, in 70% of cases, higher in the new variation.

5. Discussion

In summary, we found that Generative AI (specifically ChatGPT-4) can potentially augment a traditional Delphi study toward a hybrid expert panel. The moderate reliability between the group of AI experts compared to the group of human experts allows the conclusion that the AI experts can draw on the collective intelligence of their knowledge base and, hence, can bring a reliable additional perspective into the Delphi panel. We argue that this additional perspective could be used as a form of interrater reliability check, as well as for researcher triangulation.

This opportunity leads to easier access to collective expert knowledge, which would also be available to researchers with a limited budget, professional network, or reputation (which, we assume, would be required to engage renowned human experts into a time-consuming Delphi panel).

Our attempt to create exact digital clones of particular human experts in the form of a "digital twin expert" was, however, not successful. We can assume that accessing only the publicly available data of one human expert does not allow ChatGPT to extract sufficient knowledge to mimic the behavior and opinions of that particular person. The tacit knowledge of the human expert achieved through 10+ years of deliberate practice, training, and experience is apparently not easy to replicate. Since the AI can only rely on expertise that is somehow published, it has no access to any tacit knowledge of the experts. Future work should investigate additional evidence of the expert's thinking and innovativeness as captured by their creative or futuristic writing.

When looking at the qualitative analyses, a few insights stood out: First, the feedback provided by the AI experts is significantly longer and more detailed than the one provided by human experts. Many of the human experts did not take the time and effort to write lengthy explanations or even suggest new scenarios. By contrast, AI experts are happy to do so when instructed accordingly.

Most of the AI experts' feedback was more diplomatic and less opinionated. The 'social desirability bias' known for human expert panels was even larger for the AI experts. We argue that human feedback was sometimes more prone to trigger scenario iteration because it asked new questions to consider or even questioned the underlying assumptions of the scenario.

However, the scenario variations provided by the AI experts inspired new ideas for scenario iteration. While many of the newly suggested scenarios were rather obvious and simply variations of the existing ones, several unexpected new ideas could be found among them. In that sense, we argue that AI experts could act as a source of inspiration when iterating future scenarios or designing them from scratch.

In addition to the primary insights directly pertaining to our research question, our study also engendered a series of secondary findings of academic interest:

Despite a thoroughly performed prompt engineering process (including providing additional details and adjusting the ChatGPT temperature), there were no improvements regarding the average ratings of AI experts.

The aggregate of all ratings compared between all AI and human experts was always better aligned than the pairwise comparison of individual human and AI experts. This can be attributed to the power of the wisdom of the crowds – in this case, the collective intelligence of the ChatGPT knowledge base. Consequently, it is important to have as many experts as possible in a hybrid Delphi study to utilize this collective intelligence.

To reach a consensus among the experts is typically the main goal for a traditional (human) Delphi panel, but for a panel of AI experts, this is not the case. Here, we are rather interested in a diverse range of opinions and feedback to include more diverse perspectives. In our analyses, the results were quite the opposite: human experts had a higher standard deviation than AI experts. This apparent discrepancy might be difficult to dissolve.

In conclusion, we argue that the sort of hybrid Delphi study we designed can generate benefits in terms of (1) the possibility of an automatic researcher triangulation between human and AI experts and (2) in the extensive qualitative feedback that is provided by the AI experts that can inspire new ideas and further scenarios. Our particular design decisions influence findings from this study, and we expect that alternative approaches can generate additional knowledge about the benefits (and perhaps risks and costs) of integrating generative AI into the Delphi study process.

6. AI-infused Hybrid Delphi Study

Based on the results of our study, we present requirements and recommendations for a novel futurology method – the Hybrid Delphi Study – which enriches human experts through AI-infused expertise and feedback.

We argue that access to collective knowledge through Generative AI might enable us to include a larger variety of expertise in a Delphi expert panel than is usually available due to limited access to high-level experts.

Figure 4 illustrates the suggested AI-based "Hybrid Delphi Method." A moderator assembles a panel of human experts identified as relevant to the topic. Additional or missing expertise is identified and useful AI perspectives are generated (through ChatGPT-4 using the prompt template provided in Section 3.4). From the remaining steps of a typical Delphi study, we suggest that AI-infused expertise is involved in steps 3, 4, and 5.

We suggest the following aspects to consider when designing a Hybrid Delphi Study:

(1) Create a diverse panel (include AI perspective where needed).
(2) Develop scenarios with the human expert panel. The possibilities of involving AI expertise in this step still need to be explored.
(3) Let the scenarios be rated independently by human and AI experts. Enquire for explanations from both human and AI experts.
(4) Obtain qualitative feedback from both human and AI experts.
(5) Iterate the scenarios according to the feedback in up to three cycles. Provide transparency to the panel about which ratings and feedback stem from AI. Request both human and AI experts to suggest variations or new scenarios.
(6) The potential to involve AI experts in a synchronous concluding workshop still needs to be determined.

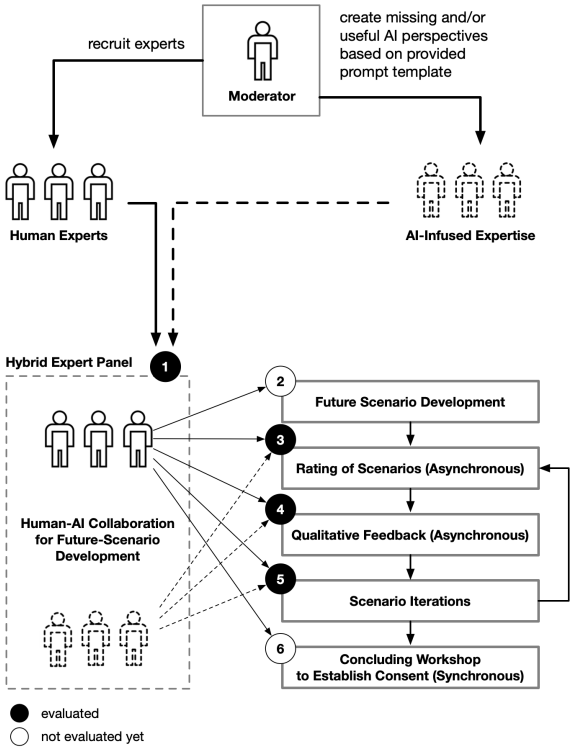

Fig. 4: Suggested AI-infused Hybrid Delphi Study.

7. Conclusions

We propose a novel futurology method – a "Hybrid Delphi Study," which includes a hybrid panel of human experts augmented by additional AI-infused expertise based on ChatGPT. In this study, we lay the foundation for such a novel approach through evaluated parts of the hybrid Delphi study.

We compared the aggregate and the pairwise reliability between scenario ratings of human and AI experts. We further explored the potential of prompt engineering towards creating a digital twin of a human expert. Additional insights were yielded by comparing the qualitative feedback provided by human and AI experts.

Overall, the findings from our study were inconclusive. Our attempt to create exact copies of a human expert in the form of a "digital twin" of a human expert was not successful. However, the moderate reliability of the aggregate ratings of the AI experts in comparison to the human experts looked promising.

We argue that this result allows us to infer the general potential of ChatGPT to expand a Delphi study toward a hybrid Delphi expert panel.

In summary, it is not yet possible to replace a human expert and his or her tacit knowledge with an AI counterpart. However, augmenting a human expert panel with AI experts seems to be a promising approach for future studies. Access to the collective intelligence of ChatGPT's knowledge base provides a valuable opportunity to infuse additional perspectives into a Delphi expert panel, allowing for additional data triangulation.

We consider our study a first step towards AI-infused Delphi studies. Further research is needed to validate our first insights and to explore the potential of future versions of ChatGPT.

The following limitations of our study have to be considered: (1) The original Delphi study with human experts was conducted in early 2021, while the AI-infused analyses were conducted in mid-2023. The last update of ChatGPT-4 was in September 2021, which resulted in a time lack of approximately six months between the human experts and the AI-infused analyses. We argue that the knowledge base available for human and AI experts is comparable, but we cannot rule out that ChatGPT might have had a small knowledge lead. (2) Our analysis covers only a part of the original Delphi study. We have not yet investigated the development of the original scenarios nor the concluding workshop. Also, we did not feed the ratings from the AI experts back to the human experts or vice versa (which we would suggest to do when conducting a full hybrid Delphi study). (3) We analyzed only selected qualitative responses. Future research will include a more extensive qualitative data analysis with independent researchers.

Finally, the following opportunities for future research have been identified: (1) The specific requirements and characteristics of the AI-human interface must be defined. (2) The potential of AI support during the scenario development shall be investigated in future work. (3) A concluding workshop, as is common practice in traditional Delphi studies, must be explicitly designed for a hybrid panel. Whether feedback from the AI experts would influence the opinions of the human experts and vice versa warrants further research.

Finally, some ethical concerns might arise from our suggested Hybrid Delphi Study. For example, it is well-known that ChatGPT tends to "hallucinate" and involve fictitious information. However, compared to factual questions where hallucinations might have a negative impact, for scenario development, this is less of a problem because we are not dealing with factual but potential realities. Another concern might relate to the transparency of AI involvement. In response to this concern, we suggest clear communication to all members about the role of AI in a hybrid expert panel.

In conclusion, we consider the presented study a step towards a novel method for future-oriented design. It explores the solution space of AI-infused future methods. As such it does not want to suggest a final answer but to open a discourse in the field of designing possible futures. The research wants to raise new questions about the role of AI in imagining and designing possible futures. The first results look promising, but substantial further research is required before the suggested AI-based hybrid Delphi method can be transformed into a valid method for future-oriented design.

8. References

Abedin, B., Meske, C., Junglas, I., Rabhi, F., & Motahari-Nezhad, H. R. (2022). Designing and Managing Human-AI Interactions. *Information Systems Frontiers*, *24*(3), 691–697. https://doi.org/10.1007/s10796-022-10313-1

Arvanitidis, G., Hansen, L. K., & Hauberg, S. (2018). Latent space oddity: On the curvature of deep generative models. *6th International Conference on Learning Representations, ICLR 2018*.

Bomström, H., Annanperä, E., Kelanti, M., Xu, Y., Mäkelä, S.-M., Immonen, M., Siirtola, P., Teern, A., Liukkunen, K., & Päivärinta, T. (2022). Digital Twins About Humans—Design Objectives From Three Projects. *Journal of Computing and Information Science in Engineering*, *22*(5). https://doi.org/10.1115/1.4054270

Bruynseels, K., Santoni de Sio, F., & van den Hoven, J. (2018). Digital Twins in Health Care: Ethical Implications of an Emerging Engineering Paradigm. *Frontiers in Genetics*, *9*, 31. https://doi.org/10.3389/fgene.2018.00031

Carmel, E., Avital, M., Gray, P., Kallinikos, J., & King, J. L. (2011). Teaching Foresight and the Future. In M. Chiasson, O. Henfridsson, H. Karsten, & J. I. DeGross (Eds.), *Researching the Future in Information Systems* (pp. 291–293). Springer.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., … Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Elshan, E., Siemon, D., De Vreede, T., De Vreede, G.-J., Oeste-Reiß, S., & Ebel, P. (2022). Requirements for AI-

based Teammates: A Qualitative Inquiry in the Context of Creative Workshops. Proceedings of the 55th Hawaii International Conference on System Sciences (HICSS).
Ericsson, K. A. (2003). The Acquisition of Expert Performance as Problem Solving: Construction and Modification of Mediating Mechanisms through Deliberate Practice. In J. E. Davidson & R. J. Sternberg (Eds.), *The Psychology of Problem Solving* (pp. 31–84). Cambridge University Press. https://doi.org/10.1017/CBO9780511615771.003
Grieves, M. W. (2019). Virtually intelligent product systems: Digital and physical twins. In S. Flumerfelt , K. Schwartz, D. Mavris & S. Briceno (Eds.) *Complex Systems Engineering: Theory and Practice* (pp. 175-200). American Institute of Aeronautics and Astronautics.
Haghighi, S., Jasemi, M., Hessabi, S., & Zolanvari, A. (2018). PyCM: Multiclass confusion matrix library in Python. *Journal of Open Source Software*, *3*(25), 729.
Hovorka, D., & Peter, S. (2019). How the Future is Done. Proceedings of the 52nd Hawaii International Conference on System Sciences (HICSS).
Lai, Y., Kankanhalli, A., & Ong, D. (2021). *Human-AI Collaboration in Healthcare: A Review and Research Agenda*. Proceedings of the 54th Hawaii International Conference on System Sciences (HICSS). http://hdl.handle.net/10125/70657
Leimeister, J. M. (2010). Collective intelligence. *Business & Information Systems Engineering*, *2*(4), 245–248.
Linstone, H. A., & Turoff, M. (1975). The Delphi method. Addison-Wesley Reading, MA.
Ma, Y., Liu, J., Yi, F., Cheng, Q., Huang, Y., Lu, W., & Liu, X. (2023). AI vs. Human—Differentiation Analysis of Scientific Content Generation (arXiv:2301.10416). arXiv.
OpenAI. (2023). GPT-4 and ChatGPT Overview. https://openai.com/research/gpt-4
Sablatzky, T. (2022). The Delphi Method. Hypothesis: Research Journal for Health Information Professionals, 34(1). https://doi.org/10.18060/26224
Seeber, I., Waizenegger, L., Seidel, S., Morana, S., Benbasat, I., & Lowry, P. B. (2020). Collaborating with technology-based autonomous agents: Issues and research opportunities. Internet Research, 30(1), 1–18. https://doi.org/10.1108/INTR-12-2019-0503
Shell International BV. (2008). Scenarios: An Explorer's Guide. Shell.
Siemon, D., Elshan, E., de Vreede, T., Oeste-Reiß, S., de Vreede, G.-J., & Ebel, P. (2022, November 7). Examining the Antecedents of Creative Collaboration with an AI Teammate. International Conference on Information Systems (ICIS).
Surowiecki, J. (2004). The wisdom of crowds. Doubleday.
Susskind, R. E., & Susskind, D. (2022). The future of the professions: How technology will transform the work of human experts (Updated edition). Oxford University Press.
Thoring, K., Huettemann, S., & Mueller, R. M. (2023). The Augmented Designer: A Research Agenda for Generative AI-Enabled Design. Proceedings of the Design Society, 3, 3345–3354. https://doi.org/10.1017/pds.2023.335
Thoring, K., Klöckner, H. W., & Mueller, R. M. (2022). Designing the Future With the "Delphi Design Sprint": Introducing a Novel Method for Design Science Research. Proceedings of the 55th Hawaii International Conference on System Sciences (HICSS).
Thoring, K., Mueller, R. M., & Klöckner, H. W. (2023). Mind the Future Gap: Introducing the FOD Framework for Future-Oriented Design. Proceedings of the 56th Hawaii International Conference on System Sciences (HICSS). https://hdl.handle.net/10125/103347
Vallat, R. (2018). Pingouin: Statistics in Python. Journal of Open Source Software, 3(31), 1026. https://doi.org/10.21105/joss.01026
Wang, Y., Shen, S., & Lim, B. Y. (2023). RePrompt: Automatic Prompt Editing to Refine AI-Generative Art Towards Precise Expressions. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 1–29. https://doi.org/10.1145/3544548.3581402
Wilson, H. J., & Daugherty, P. R. (2018). Collaborative Intelligence: Humans and AI Are Joining Forces. Harvard Business Review, 96(4), 114–123.