5-11-2023

HUMAN-AI COLLABORATION IN CONCEPTUALIZING DESIGN SCIENCE RESEARCH STUDIES: PERCEIVED HELPFULNESS OF GENERATIVE LANGUAGE MODEL'S SUGGESTIONS

Lucas Memmert
University of Hamburg, lucas.memmert@uni-hamburg.de

Izabel Cvetkovic
University of Hamburg, izabel.cvetkovic@uni-hamburg.de

Eva Bittner
University of Hamburg, bittner@informatik.uni-hamburg.de

HUMAN-AI COLLABORATION IN CONCEPTUALIZING DESIGN SCIENCE RESEARCH STUDIES: PERCEIVED HELPFULNESS OF GENERATIVE LANGUAGE MODEL'S SUGGESTIONS

Research Paper

Lucas Memmert, Universität Hamburg, Germany, lucas.memmert@uni-hamburg.de

Izabel Cvetkovic, Universität Hamburg, Germany, izabel.cvetkovic@uni-hamburg.de

Eva Bittner, Universität Hamburg, Germany, eva.bittner@uni-hamburg.de

Abstract

Solving complex problems was named a new challenge for research on human-AI collaboration. In our study we focus on a particular means to solving complex problems: design science research (DSR). We investigate whether AI, more specifically, generative language models (GLM), can support an individual in conceptualizing DSR studies by making helpful suggestions. To do so we use extracts of a published DSR study and have GPT-3, a GLM provide suggestions for aspects of this study. These suggestions are then evaluated in a survey (n=33) regarding their helpfulness. Results show that GLM suggestions are perceived to be helpful, with some variation depending on expertise. Reported interest in using such a tool in the future was high. Describing how GLMs can offer helpful suggestions we contribute toward a DSR tool support ecosystem and, more generally, towards knowledge on how humans and (generative) AI systems can team up to solve complex problems.

Keywords: GPT-3, language model, design science research, human-AI collaboration

1 Introduction

With advances in artificial intelligence (AI) there is an increased research interest in teaming up of AI systems and humans (Seeber et al., 2020a). A new challenge is solving complex problems through human-AI collaboration (Dellermann et al., 2019; Akata et al., 2020). In our study, we envision human-AI collaboration for a particular approach to solving complex problems: design science research (DSR).

Despite DSR's importance a lack of tool support for conducting DSR studies was highlighted, accompanied by a call for research to establish a DSR tool support ecosystem (Morana et al., 2018b). Tools have been proposed for facilitating DSR researchers in different aspects such as collaboration within the team, research process documentation or result communication (e.g., vom Brocke et al., 2017). Some tools leverage intelligent features, e.g., adding knowledge elements via a conversational agent (Gau et al., 2022). In making the vast DSR knowledge more accessible they might particularly benefit novices (Gau et al., 2022). However, these tools focus on supporting researchers on a meta-level.

We investigate how this support on meta-level can be complemented through support on a content-level. More specifically, we propose to imagine an AI-based tool which can be used by individual researchers when developing a concept for a DSR study. Throughout the DSR project, with the human incrementally building and adjusting the concept, the tool would make suggestions on different aspects of the concept to inspire the human to explore the problem and solution space (vom Brocke et al., 2020), and the in-between relationships (Venable, 2006) more comprehensively. Such suggestions would have to be

treated with caution and would have to be rigorously grounded by the researcher before inclusion into the concept. Thus, instead of providing general guidelines, e.g., on how to formulate a design principle (i.e., meta-level support), the tool would suggest an actual, study-relevant design principle, based on the information the user has already entered (i.e., content-level support). The user could generate new suggestions to receive further inspiration. Thereby, the user and the AI would iteratively work towards the joined goal (DSR concept) by adding and adjusting design components, considering each other's contributions (i.e., *collaboration*, Bedwell et al., 2012). To aid researchers, the tool should provide suggestions for the specific study, inspiring them to enhance the concept. As DSR studies are concerned with novel problems or solutions to make a contribution (Gregor and Hevner, 2013), creating a pool of potential suggestions beforehand for all the studies different user might want to conduct seems difficult.

Instead of creating a pool of potential suggestions for the AI to select from, we use a *generative language model* (GLM). GLM have been shown to be able to *generate* context-specific text based on a given input (Brown et al., 2020), eliminating the problem of having to create a pool of potential suggestions beforehand. In this study we investigate how suggestions generated by 'GPT-3' (Brown et al., 2020), a state-of-the-art language model, are perceived, seeking to answer the following research question: *RQ 1: To what extent are GLM suggestions perceived as helpful for conceptualizing DSR studies?*

Similarly to Gau et al. (2022) we assume that particularly novices might benefit from DSR tool support. From complex problem solving literature (psychology) it is known that novices find it more difficult to explore the problem and solution space appropriately and to build viable internal representations due to their lack of prior knowledge (Fischer et al., 2012). They therefore might find AI suggestions to be more helpful. As perceived helpfulness affects the acceptance of AI and the potential for collaboration success, we investigate the existence of individual differences in the perception of AI suggestions' helpfulness based on level of expertise. We thereby answer the following research question: *RQ2: How does the level of expertise influence the perceived helpfulness?*

To answer these research questions we use the GLM 'GPT-3' (Brown et al., 2020) to create AI suggestions. We then survey humans with different levels of expertise on their perception of the helpfulness of those AI suggestions. The results of our analysis show that the overall perception of the AI suggestions is positive. Significant differences in the perceived helpfulness based on expertise are sparse. Opposite from what we expected, higher levels of expertise occur with higher perceived helpfulness ratings. With our paper we contribute to addressing the call for research on a DSR tool ecosystem (Morana et al., 2018b), demonstrating that the existing facilitative, meta-level support can be complemented by content-level support. With our innovative approach of generating helpful AI suggestions along a canvas-like structure, we show a potential pathway for humans and AI systems, more specifically GLMs, to collaboratively design solutions for complex problems, an important research challenge (Dellermann et al., 2019).

2 Background

2.1 Design science research tool support

Solving complex problems was named a new challenge for human-AI collaboration (Dellermann et al., 2019). An approach for solving complex problems is DSR. DSR refers to "a research paradigm in which a designer answers questions relevant to human problems via the creation of innovative artifacts, thereby contributing new knowledge to the body of scientific evidence" (Hevner et al., 2010). Such innovative artifacts might include, e.g., software prototypes. A lot of literature is available including overarching DSR approaches, types of theoretical contributions through DSR or guidelines on formulating meta requirements (e.g. Chandra et al., 2015; Gregor and Hevner, 2013; Hevner et al., 2004). To aid researchers in rigorously applying DSR, tool support was suggested (Morana et al., 2018b). Proposed solutions focus on facilitating researchers in conducting DSR studies including aspects like team member collaboration, improvement of documentation, or context-independent literature suggestions (e.g., Morana et al., 2018a) and even included conversational agents to support researchers in documenting studies (e.g., Gau et al., 2022). In summary, the tools proposed facilitate researchers on a

meta-level, e.g., by providing a template for a "good" way of formulating meta requirements, or suggesting established approaches for conducting DSR studies, sometimes even by leveraging AI.

Here, we propose a complementary approach to support researchers on a second layer: on content-level (see table 1). Instead of offering information of how to do something in general, e.g., on how to formulate a meta requirement, a tool should suggest actual contents like issues, design requirement, design principles, or kernel theories for the study at hand, enhancing the concept by inspiring the researcher. Our approach uses a GLM to offer tailored suggestions based on the specific study's content.

Level	Explanation	Illustrative examples of support for the exemplary study: "Develop a design for a conversational agent to facilitate contributors in generating elaborate ideas on idea platforms"	Tool examples
Meta-level	Support independent of the specific study	• List DSR approaches, e.g., three cycle approach by (Hevner, 2007), DSR process model by (Peffers et al., 2007) • List reference literature, e.g., for phrasing for design requirements as proposed by (Chandra et al., 2015)	• MyDesignProcess.com (vom Brocke et al., 2017) • DScaffolding (Contell et al., 2017)
Content-level	Support for the specific study at hand	Inspirational suggestions for design requirements, e.g., • Conversational agents should elicit ideas by asking open-ended questions. • Conversational agents should encourage elaboration of ideas.	<*Our tool*>

Table 1. DSR tool support overview: meta-level vs. content-level support

2.2 Generative language models

GLM are large language models trained on a large corpus of text to generate free form text. They are trained with the goal to predict the next word given a certain input, i.e., given a certain context. Though they were trained task-agnostic only with this general goal, GLM have been shown to perform well on typical natural language processing tasks like translation or question answering and even on generating news articles based on a headline (Brown et al., 2020). They do so without task-specific training, i.e., without a task specific training data set or task-specific updates of the model (Brown et al., 2020). The interaction with the model (Brown et al., 2020) is performed via free-form *text-in, text-out*. The input might, e.g., consist of an instruction and/or exemplary pairs of inputs and expected outputs (e.g., 'Italy -> Rome'), followed by another input ('France ->'), for which the model then generates an output (Paris) accordingly, potentially "recognizing" the presence of pairs of countries and their capitals from the corpus used for training. The exemplary input-output-pairs are referred to as *demonstrations.* They can be distinguished from *training examples* as *demonstrations* do not lead to model weight updates, i.e., are used as additional input during inference time, whereas *training examples* typically are used to fine-tune models for a specific task including model (weight) updates (Brown et al., 2020). Depending on the number of demonstrations provided during inference time one can distinguish zero-, one- or few-shot usage (0, 1 or multiple demonstrations respectively) (Brown et al., 2020). Performance differences between the approaches vary by task (Brown et al., 2020), but even without demonstrations good results can be achieved. As the AI system's output is conditioned on the input provided during inference time (and inference parameters, e.g., randomness), the input needs to be crafted carefully. Techniques to shape the input have been developed as part of *prompt engineering* research (e.g., Mishra et al., 2020).

The utility of humans collaborating with generative AI systems was demonstrated in a variety of creative contexts (Memmert and Bittner, 2022) including joint music composition (Suh et al., 2021) or drawing (Zhang et al., 2021). Several GLM-based tools have been proposed or used to provide inspiration to humans when solving problems, e.g., an "Idea Machine", supporting humans in idea generation through theory-guided suggestions by an GLM (Di Fede et al., 2022). Gero et al. (2022) show that GLMs can provide "sparks", i.e., short sentences, described by users to offer "inspiration" and "external perspectives". Even more so, Zhu and Luo (2022) report that "GPT can perform conceptual design tasks

with a reasonable level of competence" when using it develop design suggestions for actual design challenges. Given that previous literature suggests that generative AI/GLMs can support creative activities through suggestions, we investigate whether such GLM suggestion can be helpful to solve problems tackled via DSR projects which are novel, complex, open-ended and knowledge rich (Hevner et al., 2004; Morana et al., 2018b) requiring creativity for designing solutions (Hevner et al., 2010).

Content-level support might have been difficult to achieve with more traditional approaches such as recommender systems, as creating a pool of all potential suggestions for all the possible studies the different users might want to conduct (from which an AI system could then choose and recommend suggestions to the human) does not seem feasible. Thus, instead of trying to create a repository of potential AI suggestions and training a recommender system, we instead suggest to use GPT-3 (Brown et al., 2020), a state-of-the-art GLM. The GLM is trained on a large, general corpus and is therefore flexible to *generate* suggestions regarding a variety of topics based on the human input. As a result, neither for training nor for inference a predefined set of potential AI suggestions is required. Given the potential of GLM for offering helpful suggestions for creative problem solving combined with the flexibility due to the on-demand *generation* of suggestions (instead of prior preparation), we thus believe GLMs might be suited for our case of providing suggestions and thereby content-level support for DSR.

3 Methodology

Our underlying idea was to support researchers in developing new DSR concepts with the help of AI suggestions. The human would fill in some general context information and the general problem, as well as first ideas. The system would make suggestions based on those ideas, which the user would review and might or might not find helpful to further develop the concept. As our goal was to understand the helpfulness of the AI suggestions and not to evaluate an experimental tool design, we choose to evaluate these suggestions independently to remove the potential influence of the experimental design on the results. Thus, for our study, we extracted contents from an existing, published DSR paper. We then used these contents from the study as input for the GLM in order to retrieve completions, hereafter referred to as 'suggestions'. In a survey we asked participants to rate the helpfulness of these suggestions according to a measurement instrument we reused from Rhys Cox et al. (2021).

3.1 Study approach

In our approach (figure 2) we used a foundational DSR structure (step 1) to develop prompt template strings (step 2), which we then populated with the contents of an existing study (steps 3 & 4) to generate a pool of AI suggestions (step 5) from which we select suggestions (step 6) for our survey (step 7).

Step 1. As a foundation, we used an (exemplary) canvas-like DSR structure (figure 2, #1). We derived this structure from the approach of using components like issues, theories, design requirements, design principles to describe the concept's cornerstones (in different variations e.g., available in Meth et al., 2015 and Gnewuch et al., 2017). We have enriched this canvas by adding situation, general problem, and artifact class, information typically presented in the studies' manuscript. Besides providing a more comprehensive view of the study to humans, this also allowed us to capture the studies broader context, enriching the information, which can be provided to the AI system.

Step 2. To receive context-dependent suggestions by the AI systems it needs to be provided with appropriate (context-dependent) input. As we sought to develop a re-producible and re-usable approach that would work across study contexts, we developed a fixed structure, i.e., prompt templates, which could then be populated with study-specific contents. The template strings contained structuring elements and placeholders for the specific contents of the study as well as a task description. As concepts like "design requirements" or "design principles" carry different meanings depending on context, we included DSR specific explanations according to (Meth et al., 2015).

For designing the template strings, we used prompt design techniques to improve suggestion quality. In particular we used "decomposition reframing" and "itemization reframing" (Mishra et al., 2020). "Decomposition reframing" follows the idea of splitting a larger, more complex tasks into multiple sub-

tasks. We applied this technique by identifying the individual design components for which we sought suggestions. Thus, we had suggestions for four smaller prompts (one for each design component) instead of having one large prompt to receive suggestions for the entire study concept. Additionally, we used "itemization reframing", e.g., instead of having a longer paragraph with the different issues, we created an enumerated list in order to "signal" to the GLM that we expect additional issues as a completion. We included placeholders for design components that might have already been created by the user to act as *demonstration examples* (Brown et al., 2020). However, besides information for the situation and general problem such initial user contributions are optional. In total, we developed four prompt templates, one for each design component. As in DSR, the design components are related or derived from each other (Gnewuch et al., 2017), we included design components of previous stages into the prompt for the next type of design components to increase internal consistency, e.g., the prompt template for design requirements included placeholders for issues and kernel theories and the template for design principles included placeholders for design requirements and kernel theories (see figure 1#2).

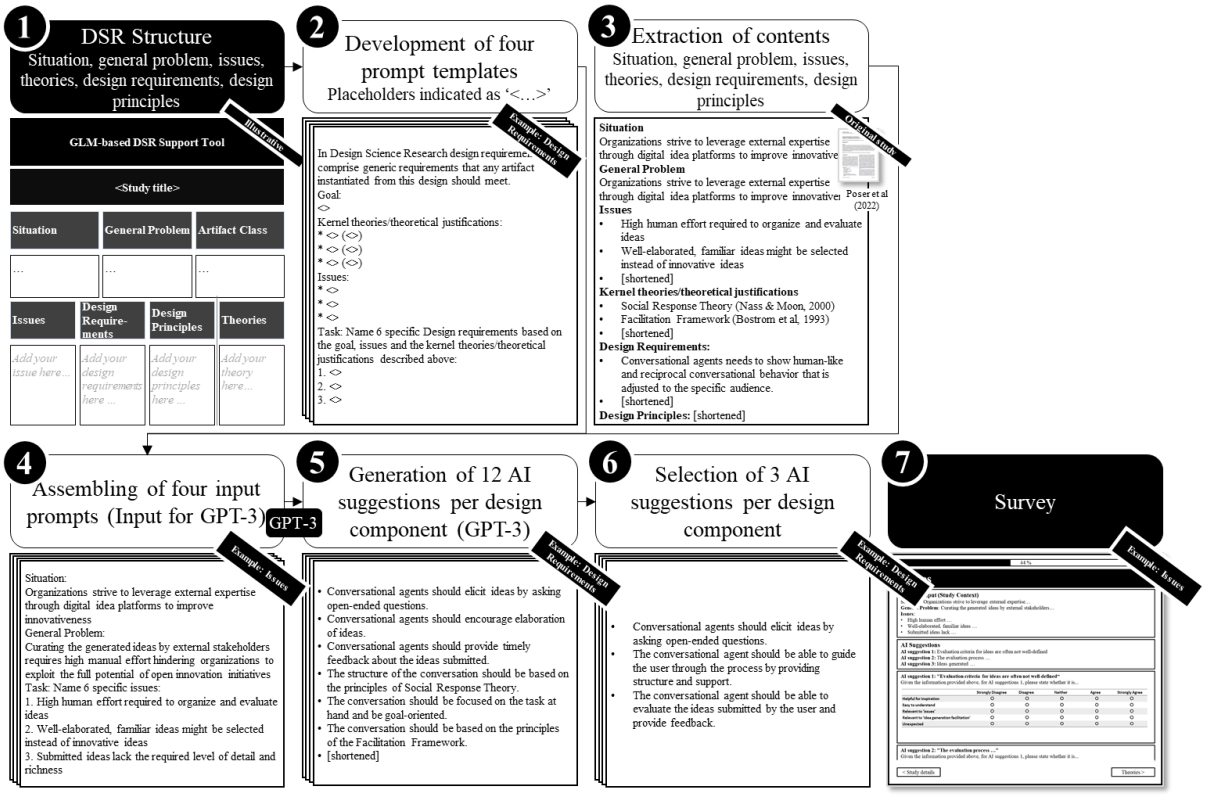

Figure 1. Study approach: from generic DSR "canvas" to context-specific survey contents with GPT-3 suggestions (steps 2-7 contain the actual data)

Step 3. We then populated our template strings with actual contents of a published DSR study. We therefore selected a *published* DSR article ("Design and Evaluation of a Conversational Agent for Facilitating Idea Generation in Organizational Innovation Processes", Poser et al., 2022) in order to ensure the foundational study idea and the resulting AI suggestions were based on a paper idea that was both interesting and relevant. Our approach, however, is general, thus any other DSR study could have been selected (or can be selected by other authors). We selected a paper published after the training data collection for the corpus for the GLM we use was completed to ensure no training data leakage occurred, i.e., to avoid the possibility of the study being part of the training data corpus. This was necessary as Carlini et al. (2023) showed that GLM might reproduce text seen during training in verbatim.

From this study, we extracted the following key conceptual elements: situation, general problem, issues, kernel theories/theoretical justifications, design requirements, and design principles. Some elements like

design requirements or design principles were clearly described as such in the original study and were thus quoted directly whereas other elements were paraphrased by the authors of this paper.

Step 4. In order to receive the study-specific suggestions we filled in the placeholders with the study contents that we had extracted from the published study.

Step 5. We then fed the resulting populated template strings as prompts into our GLM, GPT-3, to generate suggestions (see figure 1#5 for actual examples). We used the out-of-the-box GPT-3 model (i.e., without fine-tuning) at pre-defined settings (engine: text-davinci-002; temperature: 0.7).

We imagined a usage scenario as follows: while adding their ideas into the tool the user might invoke the generation of AI suggestions. The user then might check these suggestions and might update their concept, before proceeding to generate new suggestions, iteratively building their concept. For each re-generation the number of new suggestions was limited to three in order not to overwhelm the user. We assumed during typical usage a user might refresh (i.e., re-generate) the suggestions up to three times. Simulating this procedure, we generate 12 suggestions for each of the four design components (i.e., 12 suggestions per design component x 4 design component types = 48 suggestions).

Step 6. We were intrigued by the quality of some of the AI suggestions, while others seemed generic, and some occurred multiple times. Our goal was to understand if the AI system could provide helpful suggestions, not if all suggestions generated by the AI were helpful. Given our goal and the described usage scenario, we decided to select three of the twelve suggestions for each of the four design components, as even this reduced number of suggestions would allow us to understand, if suggestions were perceived to be helpful (and in which ways), but would significantly reduce the burden on participants, potentially increasing the response quality and completion rate.

Step 7. The methodology regarding the survey is described in the next subsection.

3.2 Survey

To answer our research questions regarding the helpfulness of the AI suggestion and the influence of expertise we conducted a survey gathering the relevant data and performed the corresponding analyses. Besides demographics, we asked participants to reflect their level of expertise with regard to the DSR paradigm and the topic of the study (independent variables). To determine the expertise with regard to the DSR paradigm we asked participants for how many projects or studies they had used DSR along five categories (never, 1 or 2 times, 3 to 5 times, 5 to 10 times, more than 10 times). The familiarity with regards to the topic of the foundational study (i.e., idea generation facilitation) was rated along a 5-point likert scale (‘Not at all familiar’ to ‘Extremely familiar’). Besides these more direct measures of expertise, we also measured the “general expertise” via experience in academic research, i.e., role in academia, on a 5-category scale (Bachelor student to Professor).

After introducing the foundational study (we provided the abstract of the study and asked participants to read it carefully) we asked participants to score each AI suggestions (dependent variables). To do so we had four pages (one page for every design component: issues, theories, design requirements, and design principles) with a short overview of relevant information we had extracted from the study (e.g., for the issues page we presented a short description of the situation, the general problem, and 3 issues from the study for comparison) as well as the three AI suggestions (e.g., for the issues page we presented three issues). Participants were then asked to score each AI suggestion individually (see figure 1, #7).

For measuring the helpfulness of the AI suggestions according to RQ 1 we used the instrument (both measurement dimensions and scale) from Rhys Cox et al. (2021). Similar to our study, Rhys Cox et al. (2021) used this instrument to have participants rate cues/prompts offered during ideation. According to this instrument, ratings are gathered along five dimensions: ‘helpful for inspiration’, ‘easy to understand’, ‘relevant to <task>', ‘relevant to <domain>', and ‘unexpected’. For simplicity, we refer to those five dimensions collectively as ‘helpfulness’ in this study. In our case, the <task> was to find an appropriate *design component*. The <domain> was the topic of the study, i.e., *idea generation facilitation*. For each dimension participants provide a rating on a 5-point likert scale (strongly disagree to strongly agree, coded as 1 to 5). In total participants thus provided 60 ratings with regard to the AI

suggestions: 12 suggestions (4 design components with 3 suggestions each) x 5 evaluation dimensions. At the end of the survey, we asked participants if they were interested in using or testing such an AI-based DSR support tool and if they had suggestions for further design components or features the tool should support. In total, the survey contained 23 questions and required about 15 minutes to complete.

To recruit participants with different levels of experience we reached out to students and participants of DSR university courses as well as authors who published in DSR-related tracks at information systems conferences. As we also asked addressees to forward our survey to students and colleagues the total number of recipients and consequently the response rate is unknown.

In total, we had 33 complete submissions. For the analysis we excluded partial submissions (37.1% completion rate) as most participants (75.1%) who did not complete the survey stopped before making any ratings and average ratings for participants who completed the survey were not significantly different from average ratings of participants who only partially completed it (Kruskal-Wallis: Statistic=2.292, p=0.130). Among the 33 complete submissions participants' (female=6, male=26, other=1; age: min=18, max=51, mean=30.5) levels of expertise were diverse (figure 2).

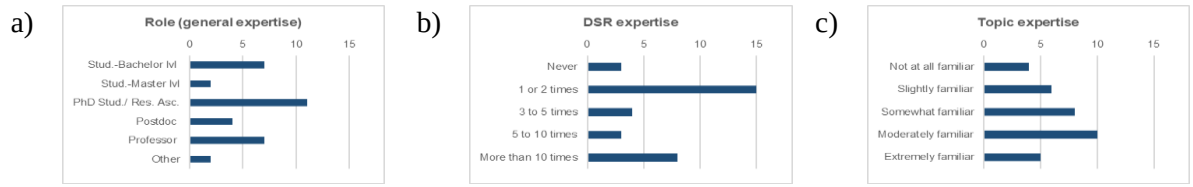

Figure 2. Expertise: a) role or general expertise, b) DSR expertise, c) topic expertise

In order to investigate the general perceived helpfulness of the AI suggestions (RQ1) we calculated the descriptive statistics of the ratings. We tested if they were significantly greater than the neutral rating and if there were any significant differences in ratings among design components, among dimensions, and among individual AI suggestions within their groups. To understand the impact of expertise on the ratings (RQ2) we performed a correlation analysis between the types of expertise and the ratings.

4 Results

We first ask whether the AI suggestions are perceived as overall helpful (RQ1). As table 2 shows, the answer is yes. On average, AI suggestions are scored at 3.5, meaning between 'neither' (3.0, i.e., neutral) and 'agree' (4.0). A Wilcoxon signed-rank test (an alternative to one sample t test for data deviating from normality, Shapiro-Wilk $p<.001$) for the alternative hypothesis 'greater than 3' revealed that ratings across all design components and dimensions except from the dimension 'unexpected' were rated significantly greater than neutral (i.e., 3.0, table 3).

In order to understand differences among design components, dimensions, and individual suggestions, we performed the Kruskal-Wallis test (non-parametric alternative to ANOVA, for data that does not fulfill the assumption of homogeneity of variances, Levene's test: $p<.001$) and corresponding Dunn's post hoc comparisons. Regarding design components (rows in table 2), pairwise post-hoc Dunn test with Bonferroni adjustments showed that AI suggestions for issues, design requirements, and design principles were rated significantly better than for theories (0.1, 0.2, and 0.3 units respectively, $p<.005$).

With regard to the dimensions (columns in table 2), AI suggestions were rated significantly worse with regard to 'unexpectedness' as compared to the other dimensions (between 0.8 and 1.1 units; $p<.001$). With regard to the individual AI suggestions there were no significant differences in rating between AI suggestions within the same design component group ($p>.05$). On a higher level, Theories show to be the least easy to understand. DRs and DPs show higher ratings on the 'Helpful for inspiration' dimension than Issues and Theories, respectively. Another indication for the AI suggestions having been perceived to be helpful was reflected in the answers to a question, which we asked after participants had seen the kinds of AI suggestions the GLM generated for our exemplary study: "Would you be interested to use a such a Design Science Research support tool that provides these kinds of AI suggestions in the future?". 81.8 % of participants responded with 'Yes', only 15.2 % answered 'No' (3.0% N/A). Looking at single design components, Issues 1 and 2 ("Evaluation criteria for ideas are often not well-defined",

"The evaluation process is often biased and subjective") display the lowest 'Unexpected' rating. The ratings for Theory suggestion 3 ("Task-technology fit") were similar across dimensions. DR2 "The conversational agent should be able to guide the user through the process by providing structure and support" displays highest rating for 'Relevant to DRs', but lowest on the dimension 'Unexpected'. DP2 "Make the conversational agent's behavior adjustable to different types of users by taking into account the user's role, knowledge and expertise, to provide an adequate level of support, and by offering different types of interactions (e.g., more directive or less directive) to users with different levels of experience." is both the most helpful one as well as the most relevant to DPs. In addition, DP2 is the only design component that positively correlates with all three expertise types on several dimensions.

Helpfulness dimensions\ AI suggestions for design components	Helpful for inspiration	Easy to understand	Relevant to <design component>	Relevant to 'idea generation facilitation'	Unexpected	Overall
Issues	3.3 (1.0)	4.0 (0.9)	3.8 (0.8)	3.5 (0.8)	2.5 (1.0)	3.4 (1.1)
1	3.4 (0.8)	4.0 (1.0)	3.8 (0.8)	3.5 (0.7)	2.3 (0.9)	3.4 (1.0)
2	3.3 (1.2)	3.9 (1.1)	3.7 (1.1)	3.4 (1.0)	2.3 (1.0)	3.3 (1.2)
3	3.2 (1.2)	4.2 (0.8)	3.9 (0.7)	3.5 (0.8)	2.8 (1.1)	3.5 (1.0)
Theories	3.4 (0.9)	3.4 (0.8)	3.5 (0.9)	3.2 (0.8)	3.1 (1.0)	3.3 (0.9)
1	3.3 (0.7)	3.1 (0.9)	3.4 (0.8)	3.1 (0.8)	3.3 (1.0)	3.3 (0.9)
2	3.5 (0.9)	3.6 (0.7)	3.5 (1.0)	3.4 (1.0)	2.9 (0.9)	3.4 (0.9)
3	3.3 (1.1)	3.6 (1.0)	3.5 (0.9)	3.2 (0.7)	3.0 (1.0)	3.3 (1.0)
Design requirements (DRs)	3.8 (0.8)	4.1 (0.9)	3.7 (1.0)	3.7 (0.9)	2.4 (0.9)	3.5 (1.1)
1	3.8 (0.8)	4.2 (0.9)	3.6 (1.1)	3.8 (0.8)	2.5 (1.0)	3.6 (1.1)
2	3.6 (0.9)	4.1 (0.9)	3.8 (1.0)	3.6 (0.9)	2.2 (0.9)	3.5 (1.1)
3	3.8 (0.8)	4.1 (0.9)	3.6 (1.0)	3.7 (0.9)	2.6 (0.8)	3.6 (1.0)
Design principles (DPs)	3.9 (0.8)	3.7 (0.9)	3.9 (0.8)	3.7 (0.7)	2.7 (1.0)	3.6 (0.9)
1	3.7 (0.9)	3.5 (1.1)	3.8 (0.9)	3.8 (0.8)	2.8 (1.0)	3.5 (1.0)
2	4.1 (0.7)	3.8 (0.8)	4.2 (0.6)	3.7 (0.6)	2.5 (0.9)	3.7 (0.9)
3	3.9 (0.7)	3.8 (0.7)	3.7 (0.8)	3.6 (0.7)	2.7 (1.0)	3.5 (0.9)
Overall	3.6 (0.9)	3.8 (0.9)	3.7 (0.9)	3.5 (0.8)	2.7 (1.0)	3.5 (0.9)

Table 2. Mean values and standard deviations for participant ratings for AI suggestions per design component (n=33)

	W	df	P
All individual ratings	842636.000	1979	< .001
Design components			
Issues	56910.000	494	< .001
Theories	41846.500	494	< .001
Design requirements	59506.500	494	< .001
Design principles	53219.000	494	< .001
Dimensions			
Helpful for inspiration	41597.000	395	< .001
Easy to understand	47366.000	395	< .001
Relevant to <design component>	43339.000	395	< .001
Relevant to 'idea generation facilitation'	31046.000	395	< .001
Unexpected	10457.000	395	1.000

Table 3. Results of Wilcoxon signed-rank test (all ratings, design components, dimensions)
Note. For the Wilcoxon sign-ranked test, the alternative hypothesis specifies that the median is greater than 3.

Influence of expertise. Regarding the influence of expertise on perceived helpfulness (RQ2) the results are heterogeneous. We analyzed correlations between the role (general expertise), topic expertise, and DSR expertise and different ratings for concepts elements averaged as well as separately for each design component. Assumption checks were performed before correlation analysis to determine the normality of the data and choose the according correlation coefficient. According to Shapiro's pairwise normality test, most of the variables did not have normal distribution. Thus, for a more comprehensive comparison,

we chose Spearman's correlation coefficient which is more robust. While most of the correlations we calculated between expertise and design components did not yield statistically significant results, some of the aspects did. In total we found 19 significant ($p<0.05$) correlations between expertise (three types) and design components groups/design components (7/12, respectively) across all five dimensions. Three design component groups correlate with role (Theories: rho=.367 (Relevant to 'idea generation facilitation'); DRs: rho=.393 (Helpfulness for inspiration), rho=.413 (Relevant to 'idea generation facilitation'); DPs: rho=.404 (Helpfulness for inspiration)). Two groups correlate with Topic Expertise (DRs: rho=.470 (Relevant to 'idea generation facilitation'); DPs: rho=.500 (Helpfulness for inspiration), rho=.387 (Relevant to design principles)). All correlations were positive, i.e., higher ratings occurred with higher expertise. Most significant correlations concerned the dimensions 'Helpfulness for inspiration' (8) and 'Relevant to idea generation facilitation' (7). We therefore performed a correlation analysis on these two dimensions and found the correlation to be positive, with strongest correlation across DRs (rho=0.7, $p<.001$).

Additional design components. In our survey participants rated AI suggestions on four design components (issues, theories, design requirements, design principles) and gained an impression of the capabilities of state-of-the-art GLM suggestions. In an open question we asked participants, which other design components they thought should be supported by such an AI tool. While one participant believed that the "core elements" were covered, other suggested additional design components like design features (as in concrete instantiations of the design principles), semantic connections or mappings between the design components to achieve a "mapping diagram", artifact type (e.g., instantiation, method, framework), types of data collection and validation, highlighting of context dependencies, and relevant literature. Most often, participants mentioned the need for evaluation-related aspects.

Additional potential features. We also asked participants which aspects in the design of a tool providing such AI suggestions they thought to be important. Participants suggested that the tool could support the formulation of design components, e.g., a transfer of 'raw' user-written requirements into design requirements of a 'correct' format, e.g., according to Gregor et al. (2020). They also suggested clustering of design knowledge to increase structure, support in the evaluation of design knowledge as well as support in selecting an appropriate prototyping framework.

Usage process. With regard to the usage process, participants had different ideas like integration into existing tools (e.g., citation manager, word processor). For integration into word processors context-depended inline suggestions ("like 'grammarly'") were mentioned. To improve the usefulness, more detailed explanation of AI suggestions as well as a summary with all suggestions were proposed. Lastly, a participant stressed the need to carefully design the usage process to mitigate potential negative effects.

5 Discussion

Overall, suggestions generated with our approach were rated positively regarding helpfulness both by less and more experienced participants, encouraging us to further explore this path of human-AI collaboration.

5.1 Perception of AI suggestions

Ratings for all design components were higher than neutral for all design components. However, ratings for theories were significantly lower than for the other design components. One interpretation is, that GLMs might be less suitable to propose such abstract ideas like theories. However, such differences need to be interpreted with caution as another explanation could be that theories are in general more difficult to interpret than, e.g., issues, and thus would receive lower ratings, independently of originating from human or AI. Future research should therefore investigate the differences also with regard to the suggestion's origin. While helpfulness ratings were positive there is still room for improvement. In our study, we used different prompt engineering techniques to increase the AI suggestion quality. However, we did not provide additional demonstration examples from other studies as part of the input prompt

during inference time, neither did we fine-tune the model. Brown et al. (2020)have shown that providing additional examples can improve the output quality significantly.

The lack of significant differences in ratings between AI suggestions within their groups might be caused by the preselection step we performed. If we had included all AI suggestions into the survey (or only generated 3 suggestions) this might have been different. However, as discussed in the methodology section, a selection of (3) relevant suggestions from the larger pool of suggestions (12) seemed to be more representative for the usage scenario, allowing us to answer the research questions while lowering the burden on survey participants by reducing the number of required ratings.

Expertise. Results regarding correlation between level of expertise and perceived helpfulness were mixed. Other than we expected novices did not perceive AI suggestions to be more helpful. On the contrary, in cases with significant correlation (particularly 'relevance to <topic>' and 'helpfulness for inspiration') these correlations indicated higher ratings with higher levels of expertise. It is likely that senior researchers or those who are more familiar with the topic might have a greater appreciation for the AI suggestions. Potentially they find the suggestions easier to mentally integrate and relate them with their existing knowledge. This might offset other effects known from CPS literature (Fischer et al., 2012). However, these assumptions need to be verified in a future study, potentially with a more fine-grained and objective assessment of the participants' expertise and a more in-depth investigation of the reasoning for the ratings.

Unexpectedness. The 'unexpected'-dimension received significantly lower ratings as compared to the remaining dimensions and was the only dimension with an average score (2.7) below neutrality (3.0). One participant reflected this aspect stating: "at no time was I really surprised by the suggestions and thought they are all things that would probably have come up in any discussions we would have as researchers/ practitioners". While collaboration among multiple humans is common in DSR projects (Sein et al., 2011), other humans might not always be cost-effectively available to have discussions. Thus, an "AI teammate" turning individual work of one human into collaborative work of a hybrid, sociotechnical ensemble, e.g., as described by Dellermann et al. (2019) might add value. Comparatively low user ratings with regards to unexpectedness are not surprising as "generative language models are trained to match the distribution of content generated by humans" (Brown et al., 2020). However, as we do not seek to replace the human by an AI, but have the AI collaborate with the human, e.g., inspiring a more comprehensive problem and solution space exploration, it might be desirable to increase unexpectedness of suggestions (see design fixation literature, e.g., Sio et al., 2015), while retaining high relevance to task and domain, i.e., design component and topic, respectively. Approaches to be explored might include providing additional examples (sourced from high-quality DSR publications) with a more appropriate level of unexpectedness (Zhu and Luo, 2022) or adjusting model parameters like temperature (i.e., level of randomness). Additionally, while with GPT-3 we used a high-performing state-of-the-art model (Brown et al., 2020), with rapid GLM advancements, our results need to be interpreted carefully and might be considered to show a lower boundary of what will be possible.

5.2 Usage of AI suggestion in a DSR support tool

In line with the overall positive ratings for the AI suggestions was the large percentage (81.8%) of participants being interested in using a tool that offers such GLM-based suggestions, with more than half of them (61.1%) being interested in testing such a prototype.

Caution when using suggestions. Though AI suggestions were rated positively they should be used with caution as part of a DSR support tool, in particular given the characteristics of GLM and the required rigor and suggested ethical behavior in (design science) research (Myers and Venable, 2014; Benke et al., 2020; Durani et al., 2021). Potential issues include lack of truthfulness of AI suggestions, lack of evidence for AI suggestions, a danger of plagiarism, a lack of understanding for the meaning, and a lack of novelty. As Lin et al. (2022) have demonstrated, GLM completions are not necessarily truthful, but might reproduce falsehoods and biases. Before including suggestion into the study concept it therefore needs to be carefully assessed to "ensure the quality of the artefact" (Myers and Venable, 2014, p. 806), which should include a reflection on ethical principles (Benke et al., 2020). Another issue

is the lack of evidence or support. DSR design components need to be thoroughly grounded in theory or practice, e.g., based on empirical findings (Goldkuhl, 2004). The AI suggestions provided by the tool, however, are not backed by empirical data or literature and therefore cannot be used directly. They might rather be treated as means for building "hypotheses" that need to be verified. A third issue is the potential danger of plagiarisms. It was shown that GLM might replicate entire passages of text from the underlying training data in verbatim based on certain prompts (Carlini et al., 2023). This, however, is usually not transparent from the output. Using the output as part of a study without proper attribution might constitute plagiarism. A fourth potential issue might be the lack of understanding for the meaning of the suggestions, i.e., the GLM cannot reason about developing the concept like a human but produces texts according to a probability distribution. It cannot make moral judgements (Floridi and Chiriatti, 2020), but these are required to be performed by the human (Durani et al., 2021). Lastly, the lack of unexpectedness we found in our study might appear to be contrary to the goal of DSR of developing *novel* artefacts. However, as Gregor and Hevner (2013, p. 344) point out, "nothing is really 'new'", everything "builds on previous ideas". Together with the previous issue (lack of understanding), this opens up an interesting research challenge: given the types of theorizing and reasoning in DSR (e.g., Kuechler and Vaishnavi, 2012; Venable, 2006) and the ability of GLMs to mimic "common sense reasoning" (Brown et al., 2020), what types of contributions (e.g., Gregor and Hevner, 2013) can we expected from GLMs?

Sociotechnical design. Given these challenges, a careful, sociotechnical design will be required, to reduce the risk of users only insufficiently investigating a problem and the solution during a DSR study, accepting even unfit AI suggestion. One participant stressed this aspect as follows: "I think you need to carefully design the whole socio-technical usage process to not design a tool that makes inexperienced DSR users to accept useless suggestions." The lack of critical engagement with AI suggestions resulting in an *overreliance* of humans on AI systems is a known problem from the related field of AI-assisted decision making (e.g., Buçinca et al., 2021). Narrow solution space exploration is also a known problem from the area of design studies referred to as *design fixation*, the phenomenon that humans can be influenced by design (aspects) they have encountered before and might as a result be less innovative (Jansson and Smith, 1991; Sio et al., 2015). In developing socio-technical solutions it might therefore be worthwhile to leverage approaches known from AI-assisted decision making and design studies literature to increase user's engagement with the AI suggestions and prevent narrow problem and solution space exploration, e.g., by having users actively request information (Buçinca et al., 2021) or by providing less common examples (Sio et al., 2015). Relevance to real-world practice might considered important for DSR (Gregor and Hevner, 2013). The sociotechnical design should include the tool's usage processes as part of a DSR project. Future research should explore how such a tool can support establishing relevance, e.g., by supporting empirical data gathering (preparation) or by summarizing already available observations from practice.

Beyond a sociotechnical design, however, it will be important to educate humans to critically reflect on AI suggestions. For our case, users would need to be educated about the limitations of such tools and that all suggestions need to be treated as unsubstantiated hypotheses, not as facts, as well as on ethical approaches towards DSR (e.g., Durani et al., 2021). More broadly, question around ownership, authorship, and attribution of AI outputs will need to be clarified (Stokel-Walker, 2023).

5.3 Towards a DSR support tool with embedded, generative AI capabilities

Generative AI has been shown to support humans in creative, open tasks (e.g., Jiang et al., 2022; Suh et al., 2021; Zhang et al., 2021). In this study, we show how GLMs can be used to support a human in solving an open, complex problem, exemplified along concept development for designing an artifact according to the DSR paradigm. Our study approach allowed us to evaluate the helpfulness of such GLMs independently from a specific tool implementation. However, future research should explore implementing a prototype (figure 3), potentially using our approach as a foundation. Instead of extracting contents from a published study, the system would use the inputs of the human from the user interface and plug these into the template strings we have developed (step 2) to generate AI suggestions,

which are then displayed to the human. With such a prototype, future studies should explore the validity across topics (select a different foundational study at step 3) or the influence of different prompt engineering techniques (adjust templates in step 2) to better understand the potentials and limitations of such an approach.

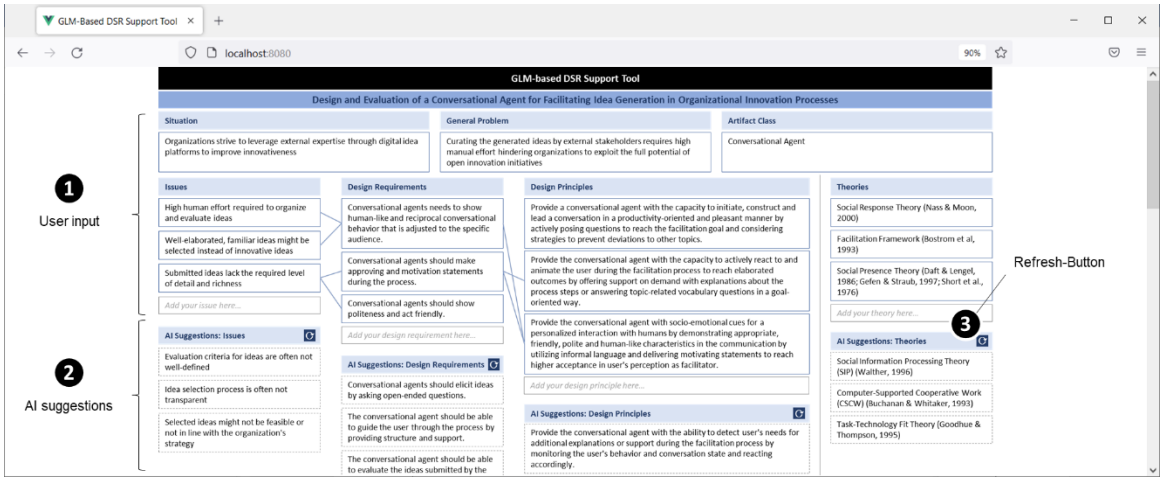

Figure 3. Prototype idea for a GLM-based DSR support tool based on our study approach: 1) user input area, 2) AI suggestion area, 3) refresh-buttons to generate new suggestions for each design component (illustrative contents for demonstration cited or paraphrased form Poser et al., 2022)

With this template-based approach, we simplify the interaction of the human with the GLM, as the inputs for the GLM (i.e., prompts) are automatically assembled in the background and the users can focus on concept development. Through this approach, generative AI is embedded into the tool and users do not have to learn to directly interact with this new technology (i.e., GLM) before working with the tool. This approach is in alignment with the challenges highlighted by Jiang et al. (2022). In their study, participants interacting with a GLM "quickly noticed that different paraphrases of the same prompt could have vastly different levels of effectiveness towards producing the desired effect". Jiang et al. (2022) state that "[…] beginners were particularly surprised when their initial zero-shot prompts did not behave as intended" and often relied on "acquiring tips and tricks". With our approach we abstract away such technology-specific aspects, allowing even non-tech-savvy users to leverage GLM capabilities while focusing on designing their artifact. Future research should explore how such human-AI collaboration affects work performance (Rafner et al., 2021), i.e., DSR concept (component) quality. More broadly, the question arises how such generative AI embedded into a tool compares to (conversational) free-text interaction with GLMs (e.g., ChatGPT). We argue that for our case, embedding generative AI could allow users to focus on concept development, without having to learn effective prompting strategies, e.g., as part of the pre-defined prompts in the tool established practices for phrasing design requirements or design principles (e.g., Chandra et al., 2015) that might be unknown to (novice) users can be included.

In this study, we focus on designing DSR solutions along a DSR canvas. However, we suggest future research to explore if other open, complex problems can be supported in a similar fashion. In alignment with our approach an established structure (figure 2, #1) like a canvas (e.g., business model canvas) could be a potential starting point for prompt template development which includes context-specific definitions of terms (#2). Such research could explore the transferability of the approach onto other complex problem solving mechanisms.

In summary, while initial AI suggestions quality is encouraging to continue with our GLM-based DSR tool support idea, more research is required to improve the suggestion quality and to carefully design of the usage process, given the characteristics of GLM and the known effects AI can have on humans' ways of working. The proposals of survey participants regarding design components, features and the usage process might be a good starting point. More broadly, future research should investigate the

applicability of our approach to other fixed structures such as established canvasses (e.g., business model canvas) to develop GLM-based tool support in a structured way.

6 Limitations

There are some limitations to our study. Only one study (of Poser et al., 2022) was included as a foundation for our survey, limiting representativeness. We decided against including more studies as the survey was already extensive. Including more studies would require participants to familiarize themselves with more, new research context(s) (incl. reading additional abstracts), increasing the burden to complete the survey. Indeed, we see that most participants who did not complete the study stopped at the first page, which included the study details, strengthening us in the trade-off we made between completion (rate) and representativeness. Even with only one study included our sample size is small.

While we used tested prompt engineering techniques, we did not provide demonstration examples from other studies. A different presentation of the study contents towards the GLM or the utilization of other prompt engineering techniques might lead to different AI suggestions, potentially affecting perceived helpfulness. Additionally, including demonstration examples can improve suggestions quality for some tasks (Brown et al., 2020). In our study, we established a first baseline. Future studies, however, should explore appropriate configurations (prompt design, inference parameters) for the task. This might include conditioning the GLM to provide more unexpected or directive suggestions. In the study at hand, the GLM reproduced the structure that was entered, i.e., if the design requirement was phrased a certain way, the GLM would try to match this structure. With more elaborate prompt design, good practices for formulating design requirements (e.g., Chandra et al., 2015; Gregor et al., 2020) could be used.

7 Conclusion and outlook

Improving our understanding of how humans and AI can team up to solve complex problems is a new research challenge (Dellermann et al., 2019). For DSR, a particular approach to solving complex problems, we demonstrated the potential of support through AI, more specifically, through GLM. With our approach, we show that by making relevant study concepts available to a GLM in a structured way, helpful suggestions can be generated. In this way, facilitative DSR tool support can be complemented by an additional layer on content-level. While human collaboration remains essential for DSR projects, our results show that GLMs might be capable of contributing during periods of individual work.

Though the general perception of the AI suggestions was positive much more research is required. In order for the tool to be helpful, AI suggestions need to be of high quality. We therefore suggest to more qualitatively explore the reasons for (dis-)satisfaction with the suggestions and the underlying individual difference. To do so, the approach described in section 4.1 should be applied to other DSR studies, potentially extended by (some) suggestions for additional design components (section 5). Future research should explore the application to other, canvas-based problem solving tasks. Additionally, effects of prompt engineering (incl. demonstration examples, Zhu and Luo, 2022), parameter tuning, and an interplay of a hybrid approach using a generative and discriminative AI might be explored.

But even then, as pointed out by a participant, the sociotechnical usage process needs to be designed carefully. Challenges like overreliance or fatigue are known from other human-AI contexts (Buçinca et al., 2021; Seeber et al., 2020b). The potential negative impact of external cues on creativity is known from design (Jansson and Smith, 1991). Applicability of this existing design knowledge (e.g., interaction processes, interface design) for human-AI teams solving complex problems should be investigated.

To fully understand the effects of such tools it will be vital to have actual users develop actual concepts with such a tool. Do researchers adjust their concept due to GLM-based suggestions? How will the resulting DSR concepts compare to those without GLM-based tool support? The high level of interest among those who completed our survey, both to offer providing additional ideas and for using or testing a GLM-based DSR support tool, motivates us to continue exploring this path of human-AI collaboration.

Acknowledgement: This research was funded by the German Federal Ministry of Education and Research (BMBF) in the context of the project HyMeKI (reference number: 01IS20057).

References

Akata, Z., D. Balliet, M. de Rijke, F. Dignum, V. Dignum, G. Eiben, A. Fokkens, D. Grossi, K. Hindriks, H. Hoos, H. Hung, C. Jonker, C. Monz, M. Neerincx, F. Oliehoek, H. Prakken, S. Schlobach, L. van der Gaag, F. van Harmelen, H. van Hoof, B. van Riemsdijk, A. van Wynsberghe, R. Verbrugge, B. Verheij, P. Vossen and M. Welling (2020). "A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence" *Computer* 53 (8), 18–28.

Bedwell, W. L., J. L. Wildman, D. DiazGranados, M. Salazar, W. S. Kramer and E. Salas (2012). "Collaboration at work: An integrative multilevel conceptualization" *Human Resource Management Review* 22 (2), 128–145.

Benke, I., J. Feine, J. R. Venable and A. Maedche (2020). "On Implementing Ethical Principles in Design Science Research" *AIS Transactions on Human-Computer Interaction* 12 (4), 206–227.

Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever and D. Amodei (2020). "Language Models are Few-Shot Learners". In *NIPS'20: Proceedings of the 34th 2020*, pp. 1877–1901.

Buçinca, Z., M. B. Malaya and K. Z. Gajos (2021). "To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making" *Proceedings of the ACM on Human-Computer Interaction* 5 (CSCW1), 1–21.

Carlini, N., D. Ippolito, M. Jagielski, K. Lee, F. Tramer and C. Zhang (2023). "Quantifying Memorization Across Neural Language Models". In *The International Conference on Learning Representations*.

Chandra, L., S. Seidel and S. Gregor (2015). "Prescriptive Knowledge in IS Research: Conceptualizing Design Principles in Terms of Materiality, Action, and Boundary Conditions". In: *2015 48th Hawaii International Conference on System Sciences*: IEEE, pp. 4039–4048.

Contell, J. P., O. Díaz and J. R. Venable (2017). "DScaffolding: A Tool to Support Learning and Conducting Design Science Research". In A. Maedche, J. vom Brocke and A. Hevner (eds.) *Designing the Digital Transformation*, pp. 441–446. Cham: Springer International Publishing.

Dellermann, D., P. Ebel, M. Söllner and J. M. Leimeister (2019). "Hybrid Intelligence" *Business & Information Systems Engineering* 61 (5), 637–643.

Di Fede, G., D. Rocchesso, S. P. Dow and S. Andolina (2022). "The Idea Machine: LLM-based Expansion, Rewriting, Combination, and Suggestion of Ideas". In: *Creativity and Cognition*. New York, NY, USA: ACM, pp. 623–627.

Durani, K., A. Eckhardt and T. and Kollmer (2021). "Towards Ethical Design Science Research". In *ICIS 2021 Proceedings*.

Fischer, A., S. Greiff and J. Funke (2012). "The Process of Solving Complex Problems" *The Journal of Problem Solving* 4 (1).

Floridi, L. and M. Chiriatti (2020). "GPT-3: Its Nature, Scope, Limits, and Consequences" *Minds and Machines* 30 (4), 681–694.

Gau, M., A. Maedche and J. vom Brocke (2022). "DSR Buddy: A Conversational Agent Supporting Design Science Research Activities" *Wirtschaftsinformatik 2022 Proceedings*.

Gero, K. I., V. Liu and L. Chilton (2022). "Sparks: Inspiration for Science Writing using Language Models". In: *Designing Interactive Systems Conference*. Ed. by F. `. Mueller, S. Greuter, R. A. Khot, P. Sweetser, M. Obrist. New York, NY, USA: ACM, pp. 1002–1019.

Gnewuch, U., S. Morana and A. Maedche (2017). "Towards Designing Cooperative and Social Conversational Agents for Customer Service". In: *Proceedings of the 38th International Conference on Information Systems (ICIS)*.

Goldkuhl, G. (2004). "Design Theories in Information Systems - A Need for Multi-Grounding" *JITTA : Journal of Information Technology Theory and Application* 6 (2), 59–72.

Gregor, S. and A. R. Hevner (2013). "Positioning and Presenting Design Science Research for Maximum Impact" *MIS Quarterly* 37 (2), 337–355.

Gregor, S., L. Kruse and S. Seidel (2020). "Research Perspectives: The Anatomy of a Design Principle" *Journal of the Association for Information Systems* 21, 1622–1652.

Hevner, March, Park and Ram (2004). "Design Science in Information Systems Research" *MIS Quarterly* 28 (1), 75.

Hevner, A. (2007). "A Three Cycle View of Design Science Research" *Scandinavian Journal of Information Systems* 19 (2).

Hevner, A., S. Chatterjee, P. Gray and C. Y. Baldwin (2010). *Design research in information systems. Theory and practice.* New York, NY, Dordrecht, Heidelberg, London: Springer.

Jansson, D. G. and S. M. Smith (1991). "Design fixation" *Design Studies* 12 (1), 3–11.

Jiang, E., K. Olson, E. Toh, A. Molina, A. Donsbach, M. Terry and C. J. Cai (2022). "PromptMaker: Prompt-Based Prototyping with Large Language Models". In: *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery.

Kuechler, W. and V. Vaishnavi (2012). "A Framework for Theory Development in Design Science Research: Multiple Perspectives" *Journal of the Association for Information Systems* 13 (6), 395–423.

Lin, S., J. Hilton and O. Evans (2022). "TruthfulQA: Measuring How Models Mimic Human Falsehoods" *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Memmert, L. and E. Bittner (2022). "Complex Problem Solving through Human-AI Collaboration: Literature Review on Research Contexts". In: *Proceedings of the 55th Hawaii International Conference on System Sciences*.

Meth, H., B. Mueller and A. Maedche (2015). "Designing a Requirement Mining System" *Journal of the Association for Information Systems* 16 (9), 799–837.

Mishra, S., D. Khashabi, C. Baral, Y. Choi and H. Hajishirzi (2020). "Reframing Instructional Prompts to GPTk's Language". In *Findings of the Association for Computational Linguistics: ACL 2022*.

Morana, S., M. Scheid, M. Gau, I. Benke and J. vom Brocke (2018a). "Research Prototype: The Design Canvas in MyDesignProcess.com" *DESRIST 2018 Conference Proceedings*.

Morana, S., J. vom Brocke, A. Maedche, S. Seidel, M. T. P. Adam, U. Bub, P. Fettke, M. Gau, A. Herwix, M. T. Mullarkey, H. D. Nguyen, J. Sjöström, P. Toreini, L. Wessel and R. Winter (2018b). "Tool Support for Design Science Research—Towards a Software Ecosystem: A Report from a DESRIST 2017 Workshop" *Communications of the Association for Information Systems*, 237–256.

Myers, M. D. and J. R. Venable (2014). "A set of ethical principles for design science research in information systems" *Information & Management* 51 (6), 801–809.

Peffers, K., T. Tuunanen, M. A. Rothenberger and S. Chatterjee (2007). "A Design Science Research Methodology for Information Systems Research" *Journal of Management Information Systems* 24 (3), 45–77.

Poser, M., G. C. Küstermann, N. Tavanapour and E. A. C. Bittner (2022). "Design and Evaluation of a Conversational Agent for Facilitating Idea Generation in Organizational Innovation Processes" *Information Systems Frontiers*.

Rafner, J., D. Dellermann, A. Hjorth, D. Verasztó, C. Kampf, W. Mackay and J. Sherson (2021). "Deskilling, Upskilling, and Reskilling: a Case for Hybrid Intelligence" *Morals & Machines* 1 (2), 24–39.

Rhys Cox, S., Y. Wang, A. Abdul, C. von der Weth and B. Y. Lim (2021). "Directed Diversity: Leveraging Language Embedding Distances for Collective Creativity in Crowd Ideation". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, pp. 1–35.

Seeber, I., E. Bittner, R. O. Briggs, T. de Vreede, G.-J. de Vreede, A. Elkins, R. Maier, A. B. Merz, S. Oeste-Reiß, N. Randrup, G. Schwabe and M. Söllner (2020a). "Machines as teammates: A research agenda on AI in team collaboration" *Information & Management* 57 (2), 103174.

Seeber, I., L. Waizenegger, S. Seidel, S. Morana, I. Benbasat and P. B. Lowry (2020b). "Collaborating with technology-based autonomous agents" *Internet Research* 30 (1), 1–18.

Sein, Henfridsson, Purao, Rossi and Lindgren (2011). "Action Design Research" *MIS Quarterly* 35 (1), 37.

Sio, U. N., K. Kotovsky and J. Cagan (2015). "Fixation or inspiration? A meta-analytic review of the role of examples on design processes" *Design Studies* 39, 70–99.

Stokel-Walker, C. (2023). "ChatGPT listed as author on research papers: many scientists disapprove" *Nature* 613 (7945), 620–621.

Suh, M., E. Youngblom, M. Terry and C. J. Cai (2021). "AI as Social Glue: Uncovering the Roles of Deep Generative AI during Social Music Composition". In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. NY, USA: ACM.

Venable, J. (2006). "The Role of Theory and Theorising in Design Science Research". In: *1st International Conference on Design Science in Information Systems and Technology*, pp. 1–18.

vom Brocke, J., P. Fettke, M. Gau, C. Houy and S. Morana (2017). "Tool-Support for Design Science Research: Design Principles and Instantiation" *SSRN Electronic Journal*.

vom Brocke, J., R. Winter, A. Hevner and A. Maedche (2020). "Special Issue Editorial – Accumulation and Evolution of Design Knowledge in Design Science Research: A Journey Through Time and Space" *Journal of the Association for Information Systems* 21 (3), 520–544.

Zhang, C., C. Yao, J. Liu, Z. Zhou, W. Zhang, L. Liu, F. Ying, Y. Zhao and G. Wang (2021). "StoryDrawer: A Co-Creative Agent Supporting Children's Storytelling through Collaborative Drawing". In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. NY, USA: ACM.

Zhu, Q. and J. Luo (2022). "Generative Pre-Trained Transformer for Design Concept Generation: An Exploration" *Proceedings of the Design Society* 2, 1825–1834.