Conversational Agent as a Black Hat: Can Criticizing Improve Idea Generation?

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

Valeria Rosenberg
University of Hamburg
valeriarosenberg10@gmail.com

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

Abstract

The Ideate phase of Design Thinking is the source of many idea creations. In this context, criticism is considered a creativity killer, yet recent studies show that criticism can be beneficial. An example of this is the black hat of one creativity method: Six Thinking Hats. It points out the weaknesses of an idea so that they are eliminated by further refining. Previous research shows that conversational agents have an advantage over humans when criticizing because of their perceived neutrality. To investigate this, we developed and implemented a conversational agent and evaluated it using an A/B test. The results of the study show that the prototype is perceived as less neutral when it criticizes. Criticizing by the conversational agent can lead to higher quality ideas. This work contributes to a better understanding of conversational agents in the black hat role as well as of their neutrality.

1. Introduction

Be it global warming, pollution, financial crises or pandemics: We are confronted with constant changes in society, technology, our coexistence and our cooperation. These factors present us with major challenges and complex problems that need to be overcome; problems that require innovative solutions. The knowledge of individuals is often insufficient in this regard (Brown, 2019), all the more necessary to work in heterogeneous, interdisciplinary groups (Schallmo and Lang, 2020). Especially when it comes to collaborative innovation, Design Thinking (DT) has proven its worth as an approach for interdisciplinary collaboration in creative processes (Tschimmel, 2012; Brown, 2019). But for an efficient and successful implementation of DT and for successful collaboration, facilitation is important (Oxley et al., 1996). Skills and experience of the person facilitating are decisive for the success of a creative collaboration. To relieve the facilitators, or to meet the high demand of these, (partial) tasks can be performed by Conversational Agents (CAs) - intelligent software that can understand and use natural language in written or spoken form (Bittner et al., 2021). Several studies have already demonstrated the successful use of CAs in supporting creativity (Tavanapour et al., 2019; Tavanapour and Bittner, 2018; Bittner et al., 2019b). Through developments in Machine Learning (ML) and artificial Intelligence (AI), CAs are becoming increasingly capable in processing and responding to user input (Seeber et al., 2018). However, many characteristics have not yet been explored (Bittner et al., 2021; Seeber et al., 2018). In particular, there is a need for CAs to appear more human, e.g., by responding empathically (Debowski et al., 2021), or knowing how to deal with creative context (Tavanapour et al., 2019). On the other hand, the very lack of humanity of the CA, in the form of neutrality, can be seen as a strength and advantage compared to the human moderators (Debowski et al., 2021; Tavanapour et al., 2019; Tavanapour and Bittner, 2018). In one study, a CA repeatedly prompts participants to make a decision during a discussion. While these prompts were perceived as neutral by the participants, the same prompts, coming from a moderator, were evaluated negatively. This observation shows an unexplored potential of moderating CAs in DT (Tavanapour et al., 2019). Moreover, in the context of ideation, there is a need for CAs that can contribute new impulses and inspirations and promote different directions in ideation (Bittner et al., 2021). In this context, the CA could not only provide support but also act as an equal collaboration partner (Seeber et al., 2018) and contribute its own perspectives (Debowski et al., 2021). An application scenario that includes all the aspects mentioned above to use a CA is the Six Thinking Hats method (De Bono, 1985). This is a creativity technique in which a group methodically adopts six different thinking perspectives in search of a solution to a problem. The hats of the colors blue, white, red, yellow, green and black each symbolize a direction of thinking. In order to utilize the expertise of all participants in each direction, during a discourse each participant should represent each hat,

URI: https://hdl.handle.net/10125/102651
978-0-9981331-6-4

i.e., each way of thinking and role, in turn (De Bono, 1985). For example, the blue hat is responsible for organizing and controlling all other hats. White is neutral and objective and is focused on facts. The green hat represents creativity and new ideas, and the black hat is logically negative and points out the weaknesses of an idea. Building on this method, a CA with the goal of a balanced team, could analyze a group as they work together in terms of roles or points of view in order to take a missing perspective (Debowski et al., 2021).

With this study, the so far not investigated impact of criticism by a CA in the context of an idea generation process is explored. For this purpose, a CA prototype is designed and implemented so that it can be investigated whether it can successfully criticize ideas. Successfully means that the criticism leads to rethinking and submitting new ideas or even that newly submitted ideas are qualitatively better. In the course of this, the effect of the CA's criticism will be explored with a focus on its neutrality. To achieve this goal, this work is guided by the following research questions: RQ 1. *How can the neutrality of the CA be tested and exploited using criticism towards the user?* RQ 2: *What effect does the developed CA have on the users with regard to neutrality and criticism in the idea generation process?*

The paper is structured as follows: Related Work provides insight into CAs, DT, and CAs in DT. The Method section describes the concept of the CA prototype comprising of the requirements analysis, design principles, and the dialog scenario. Thereafter, the study design is presented, and the development of the prototype is detailed. To answer the RQ 2, the results of the study are summarized, followed by their explanation, the contribution and limitations of this work, and recommendations for further research.

2. Related Work

2.1. Conversational Agents

"I propose to consider the question, 'Can machines think?'" (Turing, 2009) - with this first sentence of the 1950 essay Computing Machinery and Intelligence, Alan Turing heralds the era of AI. The "Imitation Game" described in it, known today as the Turing Test, is intended to help determine whether a machine can imitate human intelligence and answer questions in such a way that other people are convinced that it is also human (Turing, 2009). In 1966, ELIZA becomes the first chatbot to manage to give users the feeling that they are having a conversation with a human (Weizenbaum, 1966). Although chatbots have been developed since the dawn of computers, they are far from being suitable for everyday use. Only through ML and natural language processing (NLP), the development of chatbots achieves its breakthrough (Boonstra, 2021). In addition to chatbots, other terms such as virtual agents, digital agents, dialog systems, and conversational agents are also used synonymously (Bittner et al., 2019b).

NLP is a subfield of AI and aims to transform unstructured speech data into a structured format in such a way that machines can understand and respond or otherwise react to the text and the relevant information it contains. NLP can be further divided into Natural Language Understanding (NLU) and Natural Language Generation (NLG). NLU uses grammar and context to determine the intended meaning of a sentence, while NLG is concerned with the machine generation of text based on a given set of data (Kavlakoglu, 2020). Today, most people carry a CA in their smartphone (e.g., Google Assistant or Siri) or have one installed at home (e.g., Alexa from Amazon Echo) (Boonstra, 2021). Initially, companies focused on developing their own AI-driven CAs. More recently, by disclosing their bot platforms, companies such as Telegram, Slack, and Facebook are enabling external developers to create their own CAs to provide application extensions such as polling, integrations, and entertainment (Khan and Das, 2018). As the number of Internet users continues to increase, so does the generation of data that CAs can analyze and use to extract information (Khan and Das, 2018). Due to the current continuous development of ML, NLP and AI, more and more data and the still unexploited application areas of CAs, it can be concluded that the technology has both relevance and high potential.

2.2. Design Thinking

Design plays a role in various areas of work and life. Buchanan (1992) defines four basic categories: Design of symbolic and visual communication, Design of material objects, Design of activities and organized services, and Design of complex systems or environments for living, working, playing, and learning. All of these areas contain complex problems requiring solutions (Chasanidou et al., 2015). DT can be defined as a process for solving such problems (Schallmo and Lang, 2020), or as a way of thinking that, starting from such problems, can lead to "transformation, evolution and innovation, to new ways of living and new ways of doing business" (Tschimmel, 2012). Brown (2019) describes three areas that are critical on the path to innovation: viability, desirability, and feasibility. Specifically, viability refers to feasibility from the company's perspective. The perspective of the users and the associated wishes and needs are covered by desirability. Feasibility refers to feasibility from

a technological perspective. If all three areas have been considered, it improves the prospect of innovation (Brown, 2019; Chasanidou et al., 2015). There are different models for the DT process, such as the Hasso Plattner Institute (HPI) model consisting of the six phases Understand, Observe, Point of View, Ideate, Prototype, and Test (Thoring and Müller, 2011). The connections between the phases imply the possibility of proceeding iteratively and, if necessary, repeating phases that have already been gone through.

2.3. Conversational Agents in Design Thinking - With AI to better ideas

Like the Ideate phase of the HPI DT model, all other DT models also include a phase for generating ideas (Tschimmel, 2012). In this phase, ideas are derived from the previously determined needs of the users with the help of creativity techniques, then grouped and revised, and finally described and evaluated (Schallmo and Lang, 2020). How one can benefit from AI in the ideate phase can be shown by the example of a CA that ensures that elaborated ideas have a consistent structure, sufficient details, and a comprehensive description by encouraging users to add to them (Tavanapour and Bittner, 2018). This is because ideas often lack sufficient information, so that they score lower than other ideas in the subsequent evaluation due to the lack of presentation (Bittner and Shoury, 2019). The evaluation of ideas by CAs turns out to be more difficult than the support of idea generation and revision. Maher and Fisher (2012) take a first step in this direction by presenting an AI-based approach to evaluate ideas with respect to the criteria of novelty, value, and surprise. As in DT in general, the ideation process is not simply the activity of a single person, but a collaborative creation of new and innovative solutions in a group (Stockleben et al., 2017). To foster creativity in collaboration, brainstorming rules and techniques include e.g., ”be visual”, ”avoid criticism”, ”build on others’ ideas”, ”allow for unconventional ideas”, and ”focus on quantity” (Thoring and Müller, 2011). A good example of how a CA can facilitate such creative collaboration is brAInstorm, a web-based tool for collective digital brainstorming (Strohmann et al., 2017). The CA guides the participating individuals first through an individual idea generation and then through a collective idea evaluation. Additionally, the CA intervenes when one of the participants uses phrases that can hinder a brainstorming process (”killer phrases”), become rude, or talk too much (Strohmann et al., 2017). On the one hand, avoiding criticism is one of the most important rules of brainstorming with the aim of preventing evaluation apprehension (Siemon, 2022); on the other hand, it contradicts the selection process in the ideate phase, because for this an evaluation of the ideas is inevitable (Thoring and Müller, 2011). Moreover, the study by Tanaka et al. (2015), for example, shows that criticism can even lead to better quality ideas. They ensure that no evaluation apprehension is created by anonymity of the participants. Furthermore, Siemon et al. (2015) show that negative effects that can arise in test subjects through collaboration with other people, such as evaluation apprehension, do not occur when the collaboration takes place with an AI-based system instead.

3. Method

This study comprises the development, demonstration, and evaluation of a CA prototype following the principles of Design Science Research (DSR) methodology according to Peffers et al. (2007). Hereby, a technical artefact (CA prototype) is developed as an objective of a solution to a previously identified problem of scarcely investigated impact of criticism by a CA in the context of idea generation. In the demonstration phase, the artefact is used to employ criticism. By employing a survey after the demonstration, we evaluate the impact of criticism on idea generation.

3.1. Design Requirements & corresponding Design Principles

To create a concept for a CA, first, a requirements analysis is performed. Based on this, design principles are derived. In the further course, the design principles serve as orientation for the creation of the dialog scenario in order to subsequently create a basis for the development of the CA prototype. All three steps are covered in this chapter.

19 requirements were derived based on a representative literature review (Cooper, 1988). The requirements were categorized into general requirements, and criticism-specific requirements. In each of these categories, the requirements were considered in terms of goal-oriented behavior and interaction-oriented behavior according to (Tavanapour and Bittner, 2018).

To ensure goal-oriented behavior, general user-friendliness **(R1)** should be ensured through clear and easy-to-understand interaction (Bittner and Shoury, 2019). This can be realized by communicating all available functions, or by presenting commands by clickable buttons instead of free text fields (Bittner and Shoury, 2019). Furthermore, the interaction with the CA can be terminated at the request of the user and there is an option for continuous support (Bittner and Shoury, 2019). Efficient communication **(R2)** should

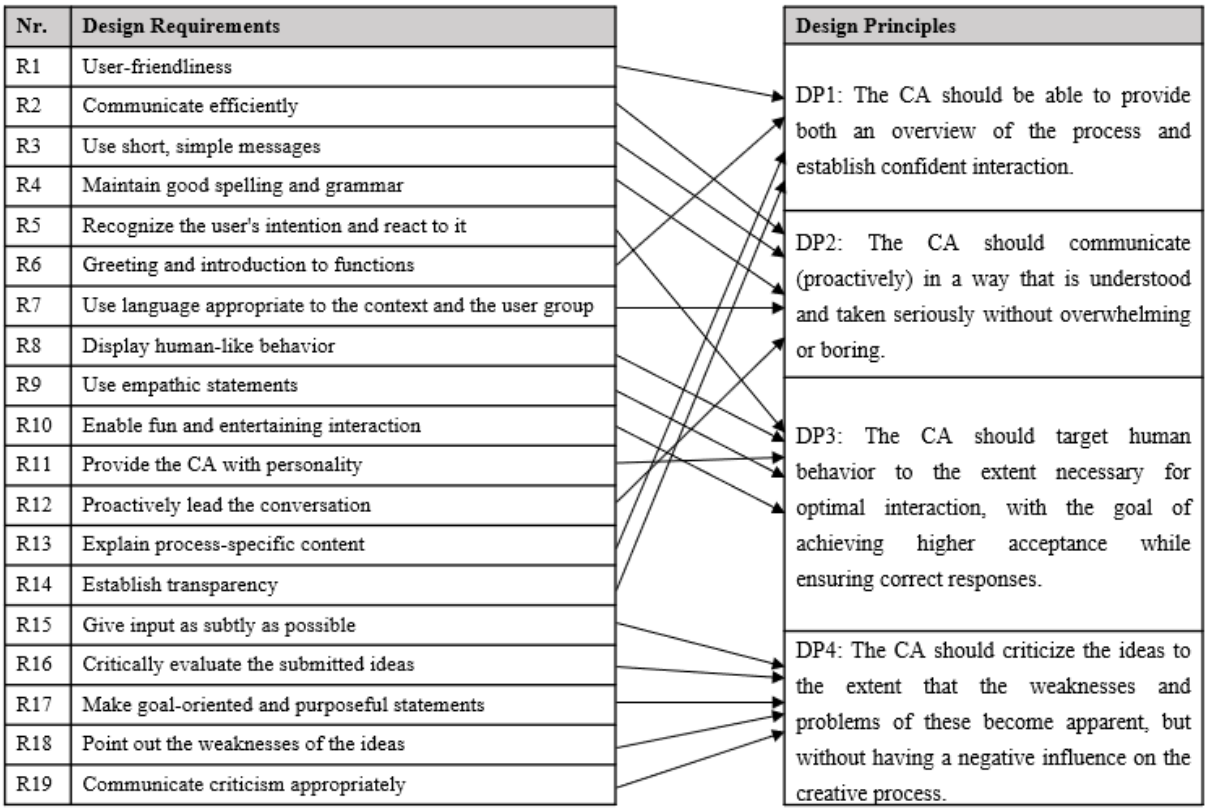

Figure 1. Design Requirements and Principles.

ensure that the needs of the user are met with minimal effort. To this end, the CA should have the shortest possible response times (Tavanapour and Bittner, 2018; Debowski et al., 2021), and help the users in as few steps as possible through optimized dialog structures (Bittner and Shoury, 2019). To make this possible, the CA could communicate the expected type of answer and check the input received from the user for correctness (Bittner and Shoury, 2019). In order not to discourage the users with long texts, short, simple messages **(R3)** should be implemented where possible (Tavanapour and Bittner, 2018). By displaying good spelling and grammar skills **(R4)**, the CA should appear as intelligent and serious as possible (Tavanapour and Bittner, 2018). The CA must be able to recognize and respond to the user's intention **(R5)** (Radziwill and Benton, 2017). If the intention of the user is not understandable, the CA should respond appropriately, e.g., by asking for more information (Bittner and Shoury, 2019). Possible misentries and dead ends should also be considered and prevented if possible (Radziwill and Benton, 2017). At the beginning of the conversation, the CA should greet the user, introduce the interaction functions **(R6)**, and explain how to interact with it (Lazarevich, 2017, 2018).

To achieve interaction-oriented behavior, the CA should use language appropriate to the context and user group **(R7)**, with an appropriate level of formality and a good vocabulary (Radziwill and Benton, 2017; Bittner et al., 2019a; Tavanapour and Bittner, 2018). Human-like behavior should be aimed for **(R8)**, such as politeness (Tavanapour and Bittner, 2018). However, the focus should not be on imitating humans, but on the successful support of the users by the CA (Radziwill and Benton, 2017). The CA should make empathic statements **(R9)**, be supportive and trusting, and respond to the users moods (Tavanapour and Bittner, 2018; Debowski et al., 2021; Radziwill and Benton, 2017). In order for the users to enjoy and be entertained by the interaction **(R10)**, the conversation with the CA should be as interesting and pleasant as possible (Tavanapour and Bittner, 2018; Radziwill and Benton, 2017). E.g., the CA should know several formulations for the same substantive statement in response to a question or use synonyms (Bittner and Shoury, 2019). With an appropriate personality of the CA **(R11)**, it should be possible to establish an emotional connection to the user (Bittner and Shoury, 2019). For example, the CA can use functions or statements personalized to the user, or to trigger emotions,

emojis can be used (Radziwill and Benton, 2017).

Regarding idea generation, the CA should not only respond to commands but proactively lead the conversation **(R12)** (Tavanapour and Bittner, 2018; Bittner and Shoury, 2019; Debowski et al., 2021), e.g., by asking relevant questions, providing conversation prompts, or initiating process-related activities (Radziwill and Benton, 2017). The CA should be able to explain both the goal of the collaboration and process-specific content **(R13)**. For example, if the CA asks the user to specify something, it should be able to explain the level of detail of the specification (Tavanapour and Bittner, 2018), or it could give examples of similar ideation tasks (Bittner et al., 2019a). To strengthen users' expectations, the CA should establish transparency **(R14)** and communicate why it is used and how interaction can improve idea generation (Bittner and Shoury, 2019; Debowski et al., 2021; Radziwill and Benton, 2017).

Regarding criticism, following requirements were identified: To ensure that the user's creativity remains in the foreground and that the creative flow is not interrupted, the CA should provide input as subtly as possible **(R15)** (Debowski et al., 2021; Bittner et al., 2021). The CA should critically question the user's ideas **(R16)**, or help the user to question the ideas themselves. In doing so, the user should be encouraged to consider the idea from different perspectives and to test the acquired information for validity (Bittner and Shoury, 2019; Wechsler et al., 2018). The CA should improve the quality of idea elaboration by making purposeful statements **(R17)** and proactively guiding the conversation (Tavanapour and Bittner, 2018). When criticism is offered, the CA should primarily point out the weaknesses of the ideas **(R18)** so that these weaknesses can be eliminated-in the spirit of the black hat role of the Six Thinking Hats method (De Bono, 1985). It must be justified in a comprehensible way why the idea is not suitable (De Bono, 1985). In order to prevent possible negative influence on the creativity of the user through criticism and, e.g., to avoid evaluation apprehension (Tanaka et al., 2015), criticism must be communicated **(R19)** appropriately (Verganti, 2016).

These requirements and corresponding principles are presented in the Fig. 1. The dialog scenario is based on the design principles and reflects all goal-oriented process steps between the CA and the user.

3.2. Study Design

To answer the RQ 1, a pilot study design is developed. The procedure of the pilot study is shown in Figure 2. First, the prototype is implemented and then a pre-test with 5 participants is conducted. In the second step, the CA is enhanced with the training data and any malfunctions that occur are corrected. In order to investigate the effect of the criticism of the CA on the ideation process, an A/B test is performed. A/B test is a controlled experiment which allows to establish causal relationships with high probability provided enough statistical power (Kohavi and Longbotham, 2017). With the improved CA, the first part of the A/B test is performed with group A. The ideas from the pre-test and the A test are used to implement the criticism function in the step 3. After the second part of the A/B test is conducted, the ideas from the two groups are combined for idea evaluation. In addition, a survey is conducted with the subjects of the A/B test.

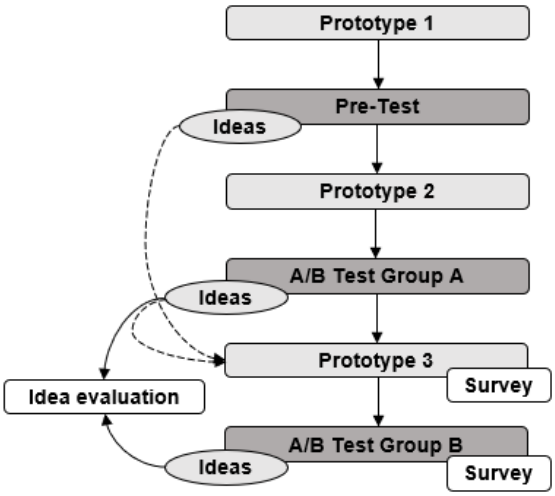

Figure 2. Study design.

The sequence of the idea generation process was based on the study design of Tanaka et al. (2015). In addition, the task for the idea generation process, the idea template, and the reference idea were also adopted from the same study. The task contains a problem for which the subjects should think of an idea for a solution. The problem is the following: "*There is a great deal of information on the Internet, including false, misleading and unsubstantiated information. What can we do to avoid or reduce the negative impact of misinformation on Internet users?*". The idea was to be formulated as a supplement to the idea template. The reference idea later served as an orientation during the expert idea evaluation in order to make the different evaluation tendencies of the idea evaluators comparable.

To conduct the A/B tests, 20 subjects, aged 20-55, were randomly recruited. They were first asked about their experience with CAs. Based on the answers, the

subjects were divided into two similarly distributed groups of ten persons each. Group A consists of one CA and ten subjects. A separate test is conducted with each subject. This test consists of an interaction with the CA. The subject is presented with a problem by the CA and is then asked to submit an idea for it by completing the idea template. After submitting the idea, the subject is given the opportunity to reconsider and change it (dummy critique). The dummy critique is used to give subjects the opportunity to change their idea so that the conversation flow with the CA is as similar as possible between Group A and Group B. Like group A, group B also consists of one CA and ten subjects. The interaction is almost identical to that in group A. The only difference is that after the idea has been submitted, the subject from group B is criticized by the CA with regard to the idea submitted. Afterwards, the subject is also given the opportunity to reconsider and change the submitted idea.

Thereupon, the experts were asked to rate all ideas (which emerged in the A/B test) in terms of novelty and practicability (scale 1-7 with 1=not at all, 7=very much), taking into account the reference idea. The experts are fairly familiar with DT and ideation through perennial work experience.

3.3. Implementation of the Criticism

The prototype CA was developed with open source NLP framework RASA (Rasa Technologies Inc, 2021). In order for the CA to be able to criticize ideas, it must be equipped with several skills. First, it must correctly understand the user's intention - to submit an idea. Then it must be able to analyze the idea and place it in the context of the problem so that it can identify possible weaknesses in the idea. Finally, it must express the criticism in a plausible and comprehensible way. This should awaken the motivation to rethink the idea and change it in such a way that the criticized weakness is eliminated. To mimic this approach for the study, the following procedure was followed: First, the problem was defined, which narrowed down the context. From the pre-test and the A/B test of group A, the submitted ideas were collected. The ideas were then grouped and categorized. Four categories emerged: Sources, Education, Algorithm, and Fallback. For the first three categories, a critique was defined to be vague enough to fit as many proposed ideas in a category as possible while being concrete enough to be perceived as plausible by the subject.

The CA was trained with the example ideas (utterances) so that it could match the ideas to one of the three categories using different keywords (entities).

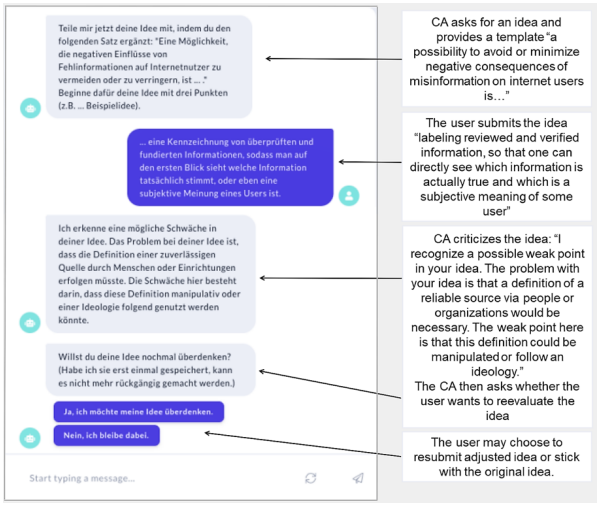

Figure 3. Snippet of the implemented critique.

In addition, slightly rephrased ideas were added to the training data to improve the CA's understanding. If the CA can categorize an idea, then it uses the critique that matches it (see example of critique to Sources in Figure 3). In case it recognizes the idea as such but cannot assign it to a category, it resorts to the Fallback category. For this purpose, a critique has been defined that is general enough to apply to most ideas related to the problem.

4. Results

4.1. User Evaluation of the prototype

Following the interaction with the CA, participants from both groups filled out a survey. The questions in the survey serve the evaluation of previously derived design principles. Apart from the answers shown in Figure 5, participants from both groups rated the CA as not boring (average rate 4,1, on a scale of 1-5, with 1=Very boring, 5=Not at all boring). Furthermore, participants felt that they knew how to communicate with the CA during the interaction (average rate 4,3, on a scale of 1-5, with 1=I felt totally lost, 5=I knew what to do). In terms of human-likeness, the average rating was 3,2 (scale 1-5, with 1=not at all human-like, 5=very human-like).

Looking at the ratings in comparison between group A and group B, we can see that the mean values show a tendency towards a different perception only for the question about neutrality.

4.2. Impact of criticism on Idea Generation

While none of the subjects in Group A revised their idea after submitting it to the CA, four out of ten in Group B did so. All ideas were then evaluated by

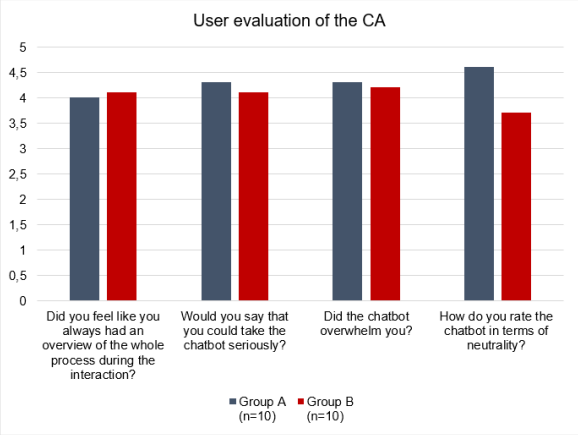

Figure 4. Answers from the post-experiment survey.

Scales left to right: 1=I felt totally lost, 5=I knew what to do; 1=I could not take it seriously at all, 5=I could take it absolutely seriously; I was totally overwhelmed, 5=I was not overwhelmed at all; 1=Not at all neutral, 5=Very neutral

innovation experts regarding novelty and practicability. The ratings of novelty and practicability presented in Figure 5 are composed of the mean value of all subjects in the respective group. The mean value was also used for the overall ratings in each case. Furthermore, a distinction is made between group B (ideas V1) and group B (ideas V2). In group B (ideas V1), the first version of the idea was included in the calculation. This means that if a subject was criticized and subsequently changed his or her idea, the change was ignored. Group B (Ideas V2), on the other hand, always included the last version of the idea. This distinction allows a comparison before and after criticism.

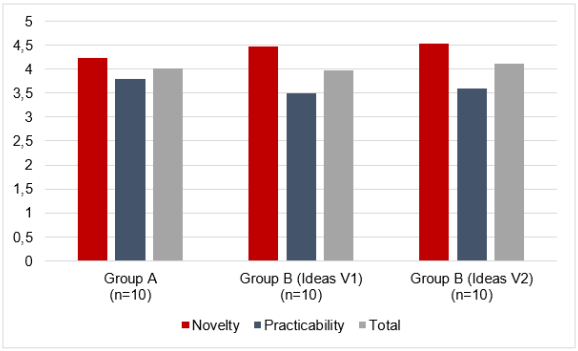

Figure 5. Expert evaluation of ideas

A comparison before and after the criticism shows that it is only after the criticism that group B tends to score better than group A in the overall evaluation. Furthermore, it can be seen that as a result of the criticism, the ratings for novelty and practicability are slightly higher, which also shows a tendency towards improvement here. Furthermore, it can be seen that all groups perform better with regard to novelty than with regard to practicability.

4.3. Impact of criticism on the perception of the CA

Sb	Evaluation of the critique (I found the critique of the chatbot to be ...)
11	not tailored to my specific idea
12	justified
13	surprisingly clever
14	interesting and constructive. However, it was somewhat amusing to be criticized by a chatbot.
15	useful, but obvious, since to filter information this always requires a party to set that filter. However, I found it interesting to be criticized.
16	appropriate. Since I could possibly not bring the topic of swarm intelligence across well enough.
17	interesting, since it was actually valid criticism of my idea.
18	valid interjection and surprisingly appropriate. In a real conversation, the criticism would have led to an interesting discussion.
19	plausible
20	surprisingly reasonable for a chatbot

Figure 6. User evaluation of the criticism

Sb = Subject

Each subject from the group B was further asked whether they recognized the criticism and whether they found the critique appropriate. All subjects recognized the critique and 9 out of 10 found it appropriate. Furthermore, subjects were asked how they perceived the criticism (question: If the chatbot criticized your idea, please complete the following sentence: ”I found the critique of the chatbot to be...”). Figure 6 summarizes the answers to this question.

If we look at the qualitative evaluation of the criticism of the test subjects for whom the criticism was evaluated as appropriate, an overview can be gained with the help of the adjectives used. The criticism is described as ”justified”, ”constructive”, ”reasonable”, ”plausible”, ”clever”, as well as twice as ”valid” and twice as ”appropriate”. Moreover, the criticism was described as ”amusing” and twice as ”interesting”. In addition, words such as ”surprising” are found twice. One time the criticism is rated as inappropriate (”not tailored to the idea”) and one time as ”obvious”.

5. Discussion

The goal of the pilot study was to find out how and if a CA can support during an ideation process by criticizing the ideas and to what extent the critique influences the ideation process, as well as the resulting perception of the subjects regarding the neutrality of the CA. The A/B test, the resulting ideas, and the feedback from the subjects collected through the survey, demonstrate an influential use of the CA prototype to support an ideation process.

The subjects' evaluation of the CA after the A/B test suggests that there is a correlation between the use of criticism and the perceived neutrality of the CA. Group B, which was criticized by the CA, rated the CA less neutral (3.7 out of 5) than Group A (4.6 out of 5), which was not criticized. Moreover, being criticized by the CA proved to be influential in terms of the idea generation process. While in group A none of the subjects changed their idea, in group B, 4 out of 10 subjects changed their idea after being criticized by the CA. With the help of the idea evaluations by the innovation experts the effect of the criticism on the quality of the ideas could be examined. While group A (without criticism) achieves a slightly better overall rating (4.02) on the quality of the ideas than group B (before criticism) (3.98), it is exactly the opposite when comparing the two groups A and B with the inclusion of criticism - group B (after criticism) (4.12) achieves a better overall rating on the quality of the ideas than group A (4.02).

5.1. User evaluation of the protoype

The implementation of design principle 1 was examined with the question "Did you feel like you always had an overview of the whole process during the interaction?" and "Did you always know how to communicate with the chatbot during the interaction?". The average score of 4.05 and 4.3 out of a maximum of 5 points, and the fact that all subjects managed to achieve the interaction goal of submitting an idea, show a successful implementation of the design principle. Design principle 2, to which the questions "How boring did you find the chatbot?", "Would you say that you could take the chatbot seriously?" and "Did the chatbot overwhelm you?" refer, could also be implemented well based on the positive evaluation. Design principle 3 is covered with the questions "How do you rate the chatbot in terms of human-likeness?" and "How do you rate the chatbot in terms of neutrality?". The fact that the CA's human-likeness was rated the worst with only 3.2 out of 5 points may be due to the fact that the CA's training data was not extensive enough. Despite its immature functionality, it is sufficient to guide the subject through the ideation process. While this does not disprove that the CA's behavior should conform to human behavior to be more acceptable (Tavanapour and Bittner, 2018), it may be indicative of the theory that passing the Turing Test may not always be the most important endeavor (Radziwill and Benton, 2017). The question about neutrality additionally serves to gather insights for RQ 2. The difference in neutrality ratings between Group A and B may indicate that there is a relationship between the use of criticism and neutrality. Overall, it appears that the concept of the prototype and the study design can be used to test how the use of criticism affects the perceived neutrality of the CA (RQ 2). The questions explained so far have been presented to both Group A and Group B and overall show a successful implementation of the design principles and thus a successful set of requirements on which the design principles are based. Since the evaluation results of the individual design principles have not reached their optimum, it can be concluded that improvement is possible through an extension in training data and functionalities.

5.2. Impact of criticism on Idea Generation

During the brainstorming process, none of the subjects from group A reconsidered or changed their idea, even though the possibility existed and the subjects were not under time pressure. For group B, 4 out of 10 subjects revised their idea. This shows that the CA's critique was good enough to convince the subjects of the possible weakness of the idea. Furthermore, it confirms the findings of Siemon et al. (2015) study that collaboration with an AI-based system does not trigger evaluation apprehension since 3 of the 4 subjects who changed their idea after the critique not only added to it but completely rethought it and submitted a new idea as a result. Moreover, it contradicts Osborn's brainstorming rule that one should avoid criticism (Osborn, cited in (Putman and Paulus, 2009)) and suggests that the subjects were not concerned about being repeatedly evaluated negatively. It is questionable whether the original purpose of the criticism, which was to identify and point out the weakness of an idea so that it could be improved by the idea creator, was fulfilled. Since the critique was adapted to the idea, but still general, it is likely that the content of the critique served as a stimulus, thereby creating inspiration for new ideas, and the comparison between group A and group B can thus be questioned, since group A lacked a possible stimulus. Based on the fact that Group B became slightly better than Group A on average only after exposure to the criticism, and the evaluation comparison of each version of the ideas - idea before criticism and changed idea after criticism - shows

possible improvements in quality, RQ 2 is answered with the fact that criticism from a CA can lead to better quality of ideas, thus confirming the hypothesis of Tanaka et al. (2015). While in their study, ideas only improve in terms of practicability, the results of this paper refute this finding and show that ideas were rated both more novel and more practical as a result of criticism. To this it must be added that some ideas became qualitatively worse after criticism and revision. This may indicate that the CA's criticism or feedback is not yet intelligent enough to intervene appropriately in such a situation and, for example, alert the subject to the deterioration of the idea. Nevertheless, the results show that the CA prototype, with more training data and more intelligent features, has the potential to represent the role of the black hat in the Six Thinking Hat method (De Bono, 1985; Debowski et al., 2021).

5.3. Impact of criticism on the perception of the CA

Group B received questions about the CA's criticism, firstly testing whether the subject perceived the criticism as such "Did the chatbot criticize your idea during the interaction?" and secondly a question about the perception of the criticism as an open question, which reads "If the chatbot criticized your idea, please complete the following sentence: "I found the critique of the chatbot to be ..."." All subjects who were criticized recognized that they were criticized and were mostly positively surprised by the CA's feedback. The adjectives used to evaluate the criticism, such as "constructive," "plausible," or "reasonable," shows a successfully implementation of the Black Hat criticism function in the sense of De Bono (1985). Some subjects rated the critique as "surprising," which could be inferred that they did not expect a CA to have this ability, or that the CA did not communicate in advance that it could criticize the ideas, which argues against the requirement that a CA should communicate its abilities transparently. This critically suggests that the overly positive perception of the CA's criticism could be due to the element of surprise. Overall, the qualitative evaluation nevertheless shows that the criticism function could be implemented expediently with only a small amount of training data. This indicates that the selection of requirements with regard to criticism and the implementation of the requirements are practicable and suitable for conducting a larger study.

6. Contributions and Limitations

The sample size was not logically derived and calculated within the scope of this work for the reliable testing of specific hypotheses or mean differences from corresponding power analyses. The A/B test was conducted with only 10 subjects randomly selected for each group. This sample size is too small to reliably compare data in statistical tests. For a detectable effect of 40 %, with statistical power of 0.8, and a significance rate of 0.05, the required sample size would be around 3000. Moreover, the sample size also does not allow for valid subgroup analyses with even smaller sizes (e.g., comparing those who change an idea vs. those who do not change an idea.). Similarly, a self-developed questionnaire was used to assess design principles - the items used here were not developed with regards to discriminatory power or other characteristics in a separate scientific process, or critically reviewed with regards to variance in response options. The data only provide indications in the sense of a pilot study and the present work can thus make an important contribution to the preparation for larger scientific investigations. Therefore, we mark this endeavour as a research-in-progress. In addition, the interaction between the subjects and the CA was not being observed. Thus, it remains open whether the subjects were distracted during the interaction, had help from others present in developing the idea, or whether the general conditions, such as the location, were equally conducive to creativity. The CA's comprehension and communication skills are limited due to limited training data. The criticism function could also be improved with more training data. In retrospect, the subject survey could have led to more insights by asking more questions. For example, when analyzing the results, it would have been interesting to know why the subjects who were criticized decided to change their idea or what prevented others from doing so.

The practical contribution of this work is an IT artifact in the form of a CA that can be used with the help of NLP to guide a user through an ideation process. It can use criticism to encourage the user to rethink an idea and improve the quality of ideas, as well as automate the ideation process as a whole. The requirements and design principles create a foundation and thus a contribution to the exploration and design of CAs in similar contexts. The dialog scenario can be reused and extended. The prototype is an example of how, despite less training data, criticism via a CA can be implemented and taken seriously by users. Furthermore, the IT artifact creates a first building block for the moderation of the Six Thinking Hats method by a CA. Thus, the work contributes to the research of CAs as facilitators and teammates. The results of the study provide preliminary evidence of how the CA's criticism function may influence perceptions regarding neutrality. Moreover, the perceived lack of human-likeness of the

CA does not seem to significantly affect the goal of the interaction.

7. Recommendations for future research

The pilot study described in this paper can be repeated with a larger sample size. In general, more training data or features can be added to the prototype to build further studies on top of it. For example, other roles of the Six Thinking Hats method can be studied. Also, it would be advisable to investigate how different the criticism of a human is perceived compared to the criticism of a CA in the same scenario and whether the CA is more accepted than a human thanks to its neutrality as assumed. Further, the prototype could be extended based on further design principles with the function that supports, motivates, inspires and informs contextually during the idea generation process. This could be combined with results from existing studies. Furthermore, it would be interesting to find out to what extent the perception of the human-likeness of a CA is related to the perception of its neutrality. Since the subjects in the study were not prepared for the CA to criticize them and this might have created a certain surprise effect that influenced the perception of the criticism, the question arises whether users perceive the criticizing differently when they know it is happening. This could be investigated with a study design similar to the one used in this paper.

References

Bittner, E. A., Küstermann, G. C., and Tratzky, C. (2019a). The facilitator is a bot: Towards a conversational agent for facilitating idea elaboration on idea platforms. In *27th European Conference on Information Systems - Information Systems for a Sharing Society, ECIS 2019*.

Bittner, E. A., Mirbabaie, M., and Morana, S. (2021). Digital facilitation assistance for collaborative, creative design processes. In *Proceedings of the Annual Hawaii International Conference on System Sciences*.

Bittner, E. A., Oeste-Reiß, S., and Leimeister, J. M. (2019b). Where is the bot in our team? toward a taxonomy of design option combinations for conversational agents in collaborative work. In *Proceedings of the Annual Hawaii International Conference on System Sciences*.

Bittner, E. A. and Shoury, O. (2019). Designing automated facilitation for design thinking: A chatbot for supporting teams in the empathy map method. In *Proceedings of the Annual Hawaii International Conference on System Sciences*.

Boonstra, L. (2021). *The Definitive Guide to Conversational AI with Dialogflow and Google Cloud*. Apress.

Brown, T. (2019). *Change by Design, Revised and Updated: How Design Thinking Transforms Organizations and Inspires Innovation*. Harper Business.

Buchanan, R. (1992). Wicked problems in design thinking. *Design Issues*, 8.

Chasanidou, D., Gasparini, A. A., and Lee, E. (2015). Design thinking methods and tools for innovation. In *Lecture Notes in Computer Science*, volume 9186.

Cooper, H. M. (1988). Organizing knowledge syntheses: A taxonomy of literature reviews. *Knowledge in Society*, 1(1):104.

De Bono, E. (1985). *Six thinking hats*. Key Porter Books.

Debowski, N., Tavanapour, N., and Bittner, E. (2021). Prototyping a conversational agent for ai-supported ideation in organizational creativity processes. In *Proceedings of the Annual Hawaii International Conference on System Sciences (HICSS)*.

Kavlakoglu, E. (2020). Nlp vs. nlu vs. nlg: the differences between three natural language processing concepts. https://ibm.co/39xlyBM, accessed on 04.2020.

Khan, R. and Das, A. (2018). Build better chatbots. *A complete guide to getting started with chatbots*.

Kohavi, R. and Longbotham, R. (2017). Online controlled experiments and a/b testing. *Encyclopedia of Machine Learning and Data Mining*.

Lazarevich, K. (2017). 7 tips for creating effective chatbot design. https://bit.ly/38V1r04, accessed on: 14.04.2022.

Lazarevich, K. (2018). How to document chatbot requirements. https://bit.ly/3wYxjdx, accessed on: 14.04.2022.

Maher, M. L. and Fisher, D. H. (2012). Using ai to evaluate creative designs. In *ICDC 2012 - 2nd International Conference on Design Creativity, Proceedings*.

Oxley, N. L., Dzindolet, M. T., and Paulus, P. B. (1996). The effects of facilitators on the performance of brainstorming groups. *Journal of Social Behavior and Personality*, 11(4):633–646.

Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems*, 24(3):45–77.

Putman, V. L. and Paulus, P. B. (2009). Brainstorming, brainstorming rules and decision making. *Journal of Creative Behavior*, 43.

Radziwill, N. M. and Benton, M. C. (2017). Evaluating quality of chatbots and intelligent conversational agents. *arXiv preprint arXiv:1704.04579*.

Rasa Technologies Inc (2021). Rasa: Open source conversational ai. https://rasa.com/, accessed on 14.06.2022.

Schallmo, D. R. and Lang, K. (2020). *Design Thinking erfolgreich anwenden*. Springer.

Seeber, I., Bittner, E., Briggs, R. O., de Vreede, G. J., de Vreede, T., Druckenmiller, D., Maier, R., Merz, A. B., Oeste-Reiß, S., Randrup, N., Schwabe, G., and Söllner, M. (2018). Machines as teammates: A collaboration research agenda. In *Proceedings of the Annual Hawaii International Conference on System Sciences (HICSS)*.

Siemon, D. (2022). Let the computer evaluate your idea: evaluation apprehension in human-computer collaboration. *Behaviour Information Technology*, pages 1–19. doi: 10.1080/0144929X.2021.2023638.

Siemon, D., Eckardt, L., and Robra-Bissantz, S. (2015). Tracking down the negative group creativity effects with the help of an artificial intelligence-like support system. In *Proceedings of the Annual Hawaii International Conference on System Sciences*.

Stockleben, B., Thayne, M., Jäminki, S., Haukijärvi, I., Mavengere, N. B., Demirbilek, M., and Ruohonen, M. (2017). Towards a framework for creative online

collaboration: A research on challenges and context. *Education and Information Technologies*, 22.

Strohmann, T., Siemon, D., and Robra-Bissantz, S. (2017). brainstorm: Intelligent assistance in group idea generation. In *Lecture Notes in Computer Science*.

Tanaka, Y., Sakamoto, Y., and Sonehara, N. (2015). The effects of criticism on creative ideation. *Cognitive Science*.

Tavanapour, N. and Bittner, E. (2018). Automated facilitation for idea platforms: Design and evaluation of a chatbot prototype. In *International Conference on Information Systems (ICIS)*.

Tavanapour, N., Poser, M., and Bittner, E. (2019). Supporting the idea generation process in citizen participation - toward an interactive system with a conversational agent as facilitator. In *Twenty-Seventh European Conference on Information Systems (ECIS)*.

Thoring, K. and Müller, R. M. (2011). Understanding design thinking: A process model based on method engineering. In *Proceedings of the 13th International Conference on Engineering and Product Design Education*.

Tschimmel, K. (2012). Design thinking as an effective toolkit for innovation. In *ISPIM Conference Proceedings*. The International Society for Professional Innovation Management (ISPIM).

Turing, A. M. (2009). Computing machinery and intelligence. In *Parsing the turing test*, pages 23–65. Springer.

Verganti, R. (2016). The innovative power of criticism. https://bit.ly/3NK1Dyc, accessed on: 14.04.2022.

Wechsler, S. M., Saiz, C., Rivas, S. F., Vendramini, C. M. M., Almeida, L. S., Mundim, M. C., and Franco, A. (2018). Creative and critical thinking: Independent or overlapping components? *Thinking Skills and Creativity*, 27:114–122.

Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.