

# Self-supervised language models in journalism: quality perception of GPT-3-written articles

Thesis in the degree programme Media, Technology and Society to obtain the  
academic degree Master of Science (M.Sc.)

submitted by

**Lasse Häufiglößner**

matriculation number: 752836

to be submitted on February 15, 2023

Supervisor: Prof. Dr. Sebastian Pranz

Co-supervisor: Prof. Dr. Melina Alexa

## Abstract

This experimental study investigates readers' perceived text quality and trust towards journalistic opinion pieces written by the language model GPT-3. GPT-3 is capable of automatically writing texts in human language and is often referred to as an artificial intelligence (AI). In a 2x2x2 within-subjects experimental design, 192 participants were presented with two randomly selected articles each for evaluation. The articles were varied with regard to the variables actual source, declared source (in each case human-written or AI-written) and the topic (1 & 2). Prior to the experimental design, participants indicated the extent to which they agreed with various statements about the trustworthiness of AI in order to capture their personal attitudes towards the topic. The study found for one, that readers considered articles written by GPT-3 to be just as good as those written by human journalists. The AI-generated versions were even rated slightly better in terms of text quality as well as the trust placed in the content. However, the effect was not statistically significant. For another no negative effect on article perception was found for texts disclosed as AI-written. Articles declared as written by an AI were mostly rated equally well or again minimally better than texts declared as human, especially regarding trust. Only the readability was rated slightly worse for the case of declaring the AI as a source. Furthermore, a correlation was found between the participants' personal attitudes towards the topic of AI and their perception of allegedly AI-written articles. For articles declared as AI-written, there are slight to moderate positive correlations of the personal attitudes towards AI with each quality rating criterion. Personal preconception thus plays a role in the perception of AI-written articles.

# Table of contents

List of figures .....	5
List of tables .....	6
1 Introduction .....	7
<i>I Factual Background</i>	
2 Natural Language Generation .....	10
2.1 Template-based implementation.....	10
2.2 Advanced implementation using artificial intelligence .....	11
3 Generative Pre-trained Transformer 3 (GPT-3).....	12
3.1 Basic operating principle.....	12
3.2 Application for text generation .....	14
3.3 Discussing potentials and skills .....	16
<i>II State of research</i>	
4 Readers' perception of computer-generated text.....	20
4.1 Influence of the actual text source.....	21
4.2 Influence of the declared text source .....	27
<i>III Empirical study</i>	
5 Research gap and open questions .....	31
6 Methodology .....	35
6.1 Experimental trials .....	35
6.2 Implementation as multi-factorial within-subjects online experiment.....	39
6.2.1 Online experiment .....	39
6.2.2 Multi-factorial within-subjects design .....	41
6.3 Operationalising of article perception .....	45
6.4 Operationalising of personal attitudes towards AI.....	48
6.5 Generation of articles.....	49
6.6 Pretest of the experimental setup.....	52
6.7 Data collection and sample.....	53
6.8 Procedure for data evaluation .....	54
7 Results.....	57
7.1 Descriptive findings .....	57
7.2 Answering research questions .....	59

*IV Concluding remarks*

8 Discussion.....	64
8.1 Implications of the results .....	64
8.2 Quality criteria and limitations .....	67
9 Conclusion .....	72
Bibliography .....	75
Appendix .....	83

## List of figures

Figure 1 Division of words into tokens by GPT-3.....	12
Figure 2 Previous knowledge regarding the term ‘artificial intelligence’ .....	57
Figure 3 Mean article ratings per actual and declared source combined.....	58
Figure 4 Mean article ratings per actual article source.....	60
Figure 5 Mean article ratings per declared article source.....	61

## List of tables

Table 1 Experimental set-up in a between-subjects-design .....	36
Table 2 Experimental set-up in a within-subjects-design.....	38
Table 3 Possible article versions as experimental treatments.....	41
Table 4 Possible experimental conditions .....	42
Table 5 Operationalisation of article perception.....	45
Table 6 Operationalisation of personal attitudes towards artificial intelligence (AI).....	48
Table 7 Group sizes for individual experimental conditions .....	55
Table 8 Correlations between personal attitudes towards AI and article perception.....	63

# 1 Introduction

The impressive progress in the field of automated language generation has attracted public attention since the release of the chatbot chatGPT in November 2022 at the latest. chatGPT is based on the third generation of OpenAI's self-learning language model 'Generative Pre-trained Transformer' (GPT-3) and has been fine-tuned for conversations (OpenAI, 2022). It is basically capable of performing any kind of linguistic content generation. Whether emails, poems, recipes, summaries, website texts or news articles, any language-processing task can be automated by the system (Floridi & Chiriatti, 2020, p. 684). The already widely discussed question of the impact of so-called artificial intelligence (AI) on society, the media and the workplace has recently moved further into focus.

Within the discipline of 'natural language generation' (NLG), which deals with the computer-based, automated production of human language (Perera & Nand, 2017, p. 1), the evolving self-learning language systems represent a new era (Dale, 2020, p. 485). Since chatGPT and GPT-3 are publicly available to everyone on the OpenAI website, such innovative artificial intelligence (AI) assistants are also of interest to the general public for the first time. Anyone can feed GPT-3 with a short task and an optional sample text to automatically develop content on any topic, in different writing styles. These new possibilities have major implications for automated content production across the entire media sphere and especially for journalism. The language automations applied in journalism so far basically use predefined sentence templates into which they insert suitable data from a database (Dale, 2020, p. 484). Thus, their application is mainly limited to basic news pieces. Simple computer-written news such as weather forecasts, financial reports or sports results have been around for a few years now (Graefe, 2016, p. 40), but have remained limited to objective and purely fact-based content. A free and arbitrary development of linguistic content in a human way is not part of the abilities, which is why no opinions, genuine argumentations or raising any new questions have been possible so far (Latar, 2015, p. 77). Logical reasoning and the formulation of sound opinions were previously considered skills unique to human authors. This could change with the emergence of self-learning language models such as GPT-3, which significantly exceed the capabilities of existing systems. Although the system has no real understanding of the world and may commit logical errors, it is capable of developing texts containing subjective views and arguments (Dale, 2021, pp. 115-116). Computer programmes capable of expressing opinions represent a new dimension in automating journalistic content. Technically, subjective content such as essays, columns or commentaries are possible. This obviously has big disruptive potential in media and journalism (Graefe et al., 2018, p. 596), which is already apparent in practice. For instance, the German newspaper 'taz' has recently started its first attempts with an AI-written

column in its online edition (Kilg et al., 2022). Given the rapid progress of such models, it is to be expected that AI-written texts will enter more and more areas of society and the public. The role that computer-generated content will play in society and the media in the future depends largely on its acceptance and quality perception by readers.

Concerning simple, fact-oriented news produced by template-based software, it is known that readers can hardly distinguish them from human versions (Graefe et al., 2018), although they are usually also perceived as less readable than human-written news. In terms of perceived credibility and journalistic expertise, they perform more or less equally to human versions (Graefe et al., 2018; Graefe & Bohlken, 2020; Jia, 2020; Wu, 2019). Regarding the performance of advanced language models like GPT-3 in journalistic tasks, little is known. Brown et al. (2020) and Clark et al. (2021), found that news pieces written by self-learning language models (GPT-3) can no longer be reliably distinguished from human versions by readers, which suggests a human-like text quality. However, a precise investigation of the quality perception of the output of GPT-3 on the basis of concrete quality criteria is still pending. This study therefore researches how journalistic opinion pieces written by GPT-3 are perceived in comparison to human versions (RQ1). It also addresses the new paradigm that subjective content is now feasible and explores the extent to which readers trust the argumentation of an AI. Moreover, the perception of quality and above all the general acceptance of computer-generated content depends not only on the pure content quality, but also on whether readers are aware of its artificial origin. In this respect, negative effects on quality perception have repeatedly been shown when articles are claimed to be computer-written (Graefe & Bohlken, 2020; Longoni et al., 2022; Tewari et al., 2021; Waddell, 2018). Consequently, it was also investigated how readers react when they know about the AI-generated origin of the articles (Graefe et al. 2018). For this purpose, the possible influence of the declared source of articles (human or AI) on reader evaluation and perception was explored (RQ2).

To shed light on this research interest, based on two opinion pieces by human journalists, an AI version of each was produced using GPT-3. Following the approach of Graefe et al. (2018), 192 German-speaking study participants were each presented two articles to evaluate in a 2x2x2 within-subjects online experiment. The article versions were varied regarding three variables, in two stages each. On the one hand, there were articles on two different topics. On the other hand, the actual source of the texts as well as the declared source in the byline were varied (in each case human-written or AI-written). This resulted in eight different article versions. One human and one AI-generated article on each of the two topics (4 versions), which in turn were provided with correct and incorrect sources (8 versions). The participants were told that they are about to assess one AI-written and one human-written article. However, the article versions were randomly assigned, while



ensuring that each participant read one article per topic, one of which was labelled as written by a human journalist and one as written by an AI. The subjects evaluated the articles using established criteria of readability, journalistic expertise and credibility (Graefe et al., 2018; Graefe & Bohlken, 2020; Sundar, 1999). In addition, the trust placed in the contributions generated was surveyed in order to take into account the new argumentation possibilities of such systems. Moreover, the relation of participants' personal attitudes towards the topic of AI and their quality perception of AI-written articles was investigated (RQ3). Before the actual experimental setting, the subjects were thus asked about their agreement with various statements on trust in AI and its capabilities as well as benevolence.

The paper first introduces the factual background (I) of the discipline of 'natural language generation' (NLG), distinguishing between the two main technical implementations. Subsequently, the functionality of self-learning language models is explained on the basis of the 'Generative Pre-trained Transformer 3' (GPT-3) which is investigated in this study. Afterwards, the practical application of the system is explained and its capabilities and quality of the output are discussed. The state of research (II) presents the previous findings on reader perception of computer-generated text as well as the influence of the source declaration (AI or human) on text perception. Based on this, the empirical part (III) first identifies research gaps and formulates research questions as well as hypotheses to be investigated. The methods section introduces the basic logic of experiments before describing the concrete implementation of the conducted online experiment. It also explains the procedure for generating the articles written by GPT-3 and the data analysis process. Subsequently, the results are presented and research questions are answered. Ultimately, the concluding remarks (IV) explain the significance of the results and place them in the state of research, as well as discussing the methodological approach in order to draw final conclusions.

## *1 Factual Background*

### 2 Natural Language Generation

The discipline of ‘natural language generation’ is the basis for all innovations made in the field of automating journalism and makes the automatic generation of content possible in the first place. It thus represents the basic technology for this paper. Therefore, the field of natural language generation will be defined and applied methods and operating principles will be explained briefly.

‘Natural language generation’ (NLG) refers to the attempt to produce meaningful text in human language (Gharat et al., 2019, p. 502). Reiter & Dale (1997) originally defined NLG as “the subfield of artificial intelligence and computational linguistics that is concerned with the construction of computer systems than can produce understandable texts in English or other human languages from some underlying non-linguistic representation of information” (Reiter & Dale, 1997, p. 1, cited in Gatt & Krahmer, 2018, p. 68). NLG technology thus systematically converts an input, mostly in the form of non-linguistic data, into an output in human language (Perera & Nand, 2017, p. 1). Available NLG software and methods offer quite a wide range of complexity and abilities (Graefe, 2016). Two main methods can be distinguished regarding the technical implementation. The most trivial method of NLG is template-driven, which inserts data into pre-existing sentence pieces (Dale, 2020, p. 484). Newer and more complex methods make use of machine-learning approaches that use self-learning mechanisms, which are often referred to as an artificial intelligence (Dale, 2020, p. 485). Accordingly, the two approaches will be referred to as ‘template-based’ and ‘AI-based’ in the following. Both approaches and their underlying principles will be explained. The capabilities and functionality of AI-based technologies are further explained in Chapter 3 on the basis of the self-learning language model GPT-3 used in this study.

#### 2.1 Template-based implementation

The most trivial method of NLG is ‘template-driven’ natural language generation. This is a relatively simple way of converting data into text, by collecting numbers from a database (Graefe, 2016) and inserting values into corresponding, predefined text templates (Dale, 2020, p. 484). The principle dates back to 1970, when simple weather reports could already be automated with this method (Glahn, 1970). For many use cases that occur in practice today, it is usually sufficient to use so-called “‘smart template’ mechanisms” (Dale, 2020, p. 484) with predefined text modules. All the necessary phrases for different use cases and topics are predefined and selected to fit the

context. Appropriate data is inserted into the gaps and, depending on the requirements, individual text components may be added or changed. Additionally, a control loop may be applied to check or refine the final results (Dale, 2020, p. 484). Furthermore, template-based systems may also be capable of analysing data to gain new insights and then incorporating them into sentences (Zydorek, 2018, p. 98). Due to their nature, they are only applicable to a limited extent in cases where simple, fact- and data-based reporting takes place (Wu, 2019, p. 3). This approach is accordingly limited to a few topics (Elkins & Chun, 2020, p. 1) and is mainly used to automate sports, financial or weather reporting (Graefe, 2016, p. 40). More complex and free text generation tasks, however, require holistic approaches that ultimately aim to truly understand human language. Such advanced language models are therefore often referred to as an artificial intelligence rather than a mere language system.

## 2.2 Advanced implementation using artificial intelligence

The previously described, simple template-based methods have been outperformed in terms of their capabilities and areas of application by newer solutions which are based on so-called artificial intelligence (Tewari et al., 2021, p. 3). In contrast to the template-based approach, such advanced systems like GPT-3 do not rely on pre-defined templates, but rather self-learning processes (Tewari et al., 2021, p. 11). Thus, it learns a language without clearly prescribed rules. Rather, GPT-3 is trained on its structure repeatedly using large amounts of language data (Brown et al., 2020, pp. 8-9) and learns to predict the statistically most likely following word based on what has been written so far (Floridi & Chiriatti, 2020, p. 684). This means that these technologies are able to automatically continue a given language input in terms of content and style. In this way, any text input can be completed. Because of the very broad range of capabilities that aim to imitate those of a human author, such systems are often referred to as an artificial intelligence (AI). The term artificial intelligence accordingly refers to computers performing tasks which are usually associated with human intelligence, such as logic reasoning, analytical thinking, problem solving or, as in the case of GPT-3, learning how human language works (Tschopp et al., 2022, p. 325). AI-driven approaches to automating language can thus be considered as part of the previously described discipline of natural language generation (NLG) (Perrotta et al., 2022, pp. 1-2). The language model GPT-3 examined in this paper makes use of the described AI-approach. The general working principles of self-learning language models will therefore be explained in the following chapter using the example of GPT-3.

### 3 Generative Pre-trained Transformer 3 (GPT-3)

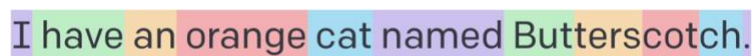
The language model used in this work is the Generative Pre-trained Transformer 3 ('GPT-3'). It is the third generation of the Generative Pre-trained Transformer language model from OpenAI. The company has been researching artificial intelligence since 2015 and develops systems that can be used for various use cases. This advanced language model uses machine learning approaches to generate natural, human-like language and is considered the most capable and largest model developed to date (Dale, 2021, p. 115). In the following, the functionality as well as the application and potential of the model are discussed.

#### 3.1 Basic operating principle

GPT-3 is a self-supervised language model, making use of deep learning techniques. Just like its predecessors GPT & GPT-2 (Brown et al., 2020, p. 8), GPT-3 uses an autoregressive language model that produces human language (Floridi & Chiriatti, 2020, p. 684). Autoregressive refers to the fact that the model independently predicts possible following language entities on the basis of already known entities (Recchia, 2021, p. 2). Thus, future values are statistically predicted based on previous values, which is called a regressive approach (Giansiracusa, 2021, p. 31; Ho, 2019). These values can represent different things (Recchia, 2021, p. 2), such as human language or computer code in the case of GPT-3 (Floridi & Chiriatti, 2020, p. 684). To process language, GPT-3 divides the entire text into individual values, so-called tokens. Tokens can be parts of words but also complete words or single characters (OpenAI, n.d.-g). Frequently occurring words are interpreted as single tokens, whereas rare words are divided into multiple tokens (see Figure 1). In English, a token is roughly equivalent to four characters or three quarters of a word (OpenAI, n.d.-b).

#### **Figure 1**

*Division of words into tokens by GPT-3*



I have an orange cat named Butterscotch.

*Note.* Taken from the OpenAI online documentation (OpenAI, n.d.-g).

In the course of autoregressive training, GPT-3 tries to find patterns in human language in enormous amounts of language data. Based on these learned patterns, predictions can then be made regarding the most possible subsequent words (Giansiracusa, 2021, p. 30). To do so, GPT-3 makes use of self-supervised learning techniques while figuring out language structures. In order

to identify structures within an existing text, it is determined how individual words or tokens depend on the characteristics of earlier text passages. This is done by automatically identifying ‘predictor variables’ (x) in a dataset of written text, which are then used to predict a ‘target variable’ (y). The variables are numerical representations of the corresponding word or token. In a text passage, all words except the last one are defined as predictor variables. The last word represents the target variable, which is the value to be forecasted out of the predictors. During the training process, complex mathematical formulas are developed that describe the relationship between predictor variables and a specific target variable and that can be used to predict the target (next word). The self-supervised approach refers to the fact that the wanted target values do not have to be manually provided by humans, but that the model can determine them on its own during the training. Just like the predictor variables, the targets are already present in the training texts. There is no need to label the data manually. This is the only way to process the enormous amounts of text on which GPT-3 was trained. In addition to the variables, various parameter values are determined and fitted into the established formula in a way that the result comes as close as possible to the actual target value, which represents the following word. The resulting algorithm consists of an equation that describes the mathematical relationship between predictors and corresponding parameters as accurately as possible for the given target value. Here, so-called deep learning techniques are applied (Floridi & Chiriatti, 2020, p. 684; Giansiracusa, 2021, pp. 30-34). This means that GPT-3 carries out its own analyses to determine the target variables and calculates new predictors not originally included in the data set, that for example represent grammatical structure, syntax or storylines (Giansiracusa, 2021, p. 33). These predictors are then put into a hierarchy (Giansiracusa, 2021, p. 33) on 96 different layers (Dale, 2021, p. 115). Put simply, it can be said that lower layers may take care of the grammar of individual sentences and syntax, while higher layers analyse complete narratives, stories, arguments and the inherent logic of an entire text. However, little can be said about the exact functioning and hierarchical interaction of the individual layers. The inner logic of GPT-3 is a black box due to the self-supervised learning without human intervention and a total number of 175 trillion applied parameters (Dale, 2021, p. 115) and can no longer be understood from the outside (Giansiracusa, 2021, p. 33).

When GPT-3 is given a text input to complete, it will return the most appropriate target value based on the predefined parameters and predictors for the particular text passage. That way, the statistically most probable following word can be obtained. That word is then again added to the input in the form of the corresponding tokens, after which the next word is determined in the same manner. All wanted words are determined on the basis of the total previous input, whereby the mechanism continues until the desired text length is reached (Giansiracusa, 2021, pp. 33-34).

The training procedure described was applied to huge datasets of written language taken from the internet and other sources. GPT-3 was trained using text datasets comprising a total of 499 billion tokens (Brown et al., 2020, p. 9). English accounts for the largest share of the dataset, although several other languages are also included (Floridi & Chiriatti, 2020, p. 684). Measured by the total amount of characters contained, 92 % of the data set originates from English, followed by French with 1.7 % and German with 1.6 % (Brown, 2020). The largest share of the total data volume is accounted for by texts from the so-called ‘common crawl’, which represents 410 billion tokens. In a crawling process, internet pages are captured and indicated in a similar way to how it is done for search engines. Thus, a large amount of text data was collected from random websites. The second largest share is made up of the ‘Books 1’ & ‘Books 2’ data sets, which together account for 67 billion tokens. Third was the ‘web text’ dataset, which consisted of specially selected, linguistically high-quality texts, with 19 billion tokens. In addition, texts from Wikipedia were included with 3 billion tokens (Brown et al., 2020, pp. 8-9).

Moreover, there are several sub-models for GPT-3 called ‘Davinci’, ‘Curie’, ‘Babbage’ and ‘Ada’, which have been trained for different tasks. However, ‘Davinci’ outperforms all other models in terms of text quality. Especially when understanding the content and intention is important, it is the most powerful model. It is also best suited for reasoning problems and causal relationships, as well as for creative writing. In addition, its version ‘text-davinci-003’ is the only model that features updated training data until June 2021, while the rest use the original training data until October 2019. As such, it is also considered as ‘GPT-3.5’. The other models also each have their own strengths, for instance they are faster and cheaper for simpler use cases than ‘Davinci’, which costs more per token used due to the higher computing capacity required (OpenAI, n.d.-d). These factors are particularly relevant for frequent users and companies, which may have faster and cheaper alternatives available depending on the use case.

### 3.2 Application for text generation

After having explained the technical implementation of GPT-3, its concrete application to generate text will be explained. GPT-3 was made available to the public in June 2020 on OpenAI’s website. It features a certain amount of text (tokens) that can be produced by anyone free of charge. Once this quota is used up, usage is charged for, making GPT-3 the first commercial product of OpenAI. Companies and other interested parties can also integrate the AI into their own products via an API and optimize it for specialised use cases (OpenAI, 2020). Part of the access is the so-called Playground, which allows anyone to use the language AI without previous experience. GPT-3 is

thus designed as a “general-purpose ‘text in, text out’ interface” (OpenAI, 2020) and can be used for any language production task by anyone. The playground, which can be used directly in the web browser features a text field, where the user can provide an input to the system. It normally consists of a prompt (formulated task) and an initial text, which is often also part of the desired output (Dale, 2021, p. 115). Ideally, the AI is both instructed to, as well as directly shown which kind of output is desired, as inserting examples helps to convey linguistic nuances (OpenAI, n.d.-g). Furthermore, some additional parameters can be defined according to preferences and use case (Dale, 2021, p. 115), which have an influence on the quality of the output. These include ‘temperature’, ‘top\_p’, ‘presence\_penalty’ and ‘frequency\_penalty’. The ‘temperature’ and ‘top\_p’ values both influence the sampling, meaning how exactly GPT-3 selects the next tokens for the output. Since the model always determines the next token or word with the highest probability by default, it can especially happen with longer texts that the output appears generic, repetitive and clumsy rather than innovative (Holtzman et al., 2020, p. 2). This is because the mathematically most probable completion does not necessarily have to be innovative or linguistically meaningful (Mann, 2019). Since the model always includes the tokens it has calculated itself so far, words that may have been incorrectly predicted are in turn be included in the prediction of the next word. The next word thus depends strongly on the previous one. This can cause accumulations of falsely predicted words and may lead to endless error loops with identical or meaningless output (Mann, 2019). Being able to choose from several of possible following tokens by defining sampling values tackles these problems. The temperature sampling value refers to the level of risk-taking when generating text. The value can range from zero to one, with a value of zero being equivalent to the standard selection of a token according to the highest probability alone (Mann, 2019). The further the value is from zero, the more the model deviates from pure selection by probability. With higher values comes more risk and variance in the output, which is suitable for creative tasks without a clear answer (OpenAI, n.d.-a). The higher the value, the more tokens with a lower probability would be suggested, which increases creativity and diversity, but also increases the risk of inappropriate answers. By contrast, low values lead to less random results. This results in fewer factual errors, but reduces creativity (Holtzman et al., 2020, p. 6). It produces deterministic and, if repeated, always the same responses. Accordingly, low temperature values are suitable for tasks with a clearly defined solution (OpenAI, n.d.-g).

The top\_p value represents an alternative selection technique to temperature. From the total number of possible token combinations, only those are selected that have a certain minimum proportion  $p$  of the total probability mass (Holtzman et al., 2020, p. 5). Thus, if top\_p is defined to be 0.2, only those token combinations are chosen that account for 20% of the total probability

set of all tokens that are considered possible (OpenAI, n.d.-a). Thus, the higher  $p$  is chosen, the smaller the selected part of the vocabulary becomes, as only a few token combinations reach the value (Holtzman et al., 2020, p. 5). OpenAI recommends changing either the temperature or the top\_p value (OpenAI, n.d.-a). In the present work, only temperature was used, as this produced the best results in the test. Top\_p did not play a role in generating the texts (see Chapter 6.5). The presence- and frequency penalty values affect the probability that the model repeats token sequences in the output. A positive presence penalty value reduces the probability of all those tokens that have already appeared in the text to appear again. As a consequence, other tokens are preferred, which makes the model more likely to pick up new words or topics. The frequency penalty works similarly, but reduces the probability of (sequences of) tokens in exact relation to how often they are already present in the text. This can prevent passages from being repeated word by word. Both penalty values work by directly changing the original probability values of the respective tokens by a certain amount. They can generally be set to values between -2 and +2. Values between 0.1 and 1 ensure a moderate reduction of repetitions, while values up to 2 strongly prevent repetitive tokens in the output. On the contrary, negative values ensure a higher probability of repetitive tokens (OpenAI, n.d.-f). The option best\_of can furthermore identify several possible completions at the same time and then select and display the best one (OpenAI, n.d.-a). However, this option was not used in this work. Finally, the max. response length defines how many tokens are generated, that is, the maximum text length of the desired output per submission.

### 3.3 Discussing potentials and skills

When discussing the capabilities of GPT-3, it becomes evident that the language model can generally be used for a very wide range of applications. Essentially, any form of text processing or generation can be done with the language model. This includes answering questions, translations, text summaries, writing emails, composing (fictional) conversations or the use as a chatbot (Floridi & Chiriatti, 2020, p. 684). Classifying items or listing examples from a certain category of things are also possible. Its logic-related skills also include arithmetic relations, the ability to play chess and even writing programming code (Dale, 2021, p. 115). The Davinci model is particularly good at creative writing, which also encompasses more complex logics, connections, causal relationships or analogies. For example, GPT-3 is able to develop fictional works and poetry (Dale, 2021, p. 117), as well as opinions and argumentation structures, which are the subject of this work.

In terms of the quality of the outputs, GPT-3 is capable of producing very human-like texts. Often the results are quite coherent, linguistically skilled and accurate (Dale, 2021, p. 115). It can write



realistic yet creative, surprising stories for a wide variety of genres (Elkins & Chun, 2020, p. 3). Thus, the model is quite capable of complex thoughts and knows how to structure arguments. It can also create quite nuanced trains of thought (Elkins & Chun, 2020, p. 12).

Despite all the praise, this does not mean that GPT-3 can be used risk-free and without human instruction or intervention. Due to some common shortcomings of the model, close supervision and control of the system is still essential, especially for complex applications. One reason for this is its general proneness to errors. This can lead to linguistically and grammatically incorrect outputs that are nonsensical or completely incomprehensible (Dale, 2021, p. 116). Rarely, non-existent words were generated during the application in the context of this work. Repetitions, sometimes of whole lines or complete paragraphs, occurred frequently. Even when setting high values for presence and frequency-penalty, which are supposed to prevent repetitions, repetitive passages occurred. Particularly in longer texts, the model then seems to get stuck on a certain thematic aspect (Dou et al., 2022, p. 25). Moreover, the entirely statistical prediction of tokens or text snippets entails some dangers to text quality. Thus, errors can also occur with regard to contents, as the outputs are not tied to facts or truth. When asking the system a question, it answers correctly in many cases, but often also incorrectly (Dale, 2021, p. 116). It can even output longer texts almost perfectly and without errors, without linguistic, logical or factual mistakes. However, this is never guaranteed. Due to the way the model works, it obviously has no actual understanding of the output content. GPT-3 uses probabilities to determine the statistically most appropriate continuation of an input text. From the system's point of view, it is irrelevant whether the generated statements are in any way meaningful or factually true (Floridi & Chiriatti, 2020, p. 684). The language model therefore only has syntactical, respectively statistical, knowledge of how words are to be arranged. It has no idea about the topical or situational context, facts or real interrelationships (Floridi & Chiriatti, 2020, pp. 688-689). Whether statements are factually and linguistically correct thus depends on whether the corresponding arrangement of the word snippets (tokens) is considered likely from the point of view of the AI, based on the training data. This problem applies to every statistical language model, which is why their output need by no means be bound to the truth (Sobieszek & Price, 2022, p. 361). Genuine basic knowledge and real human understanding cannot be assumed (Elkins & Chun, 2020, p. 12). Logical connections are also affected negatively by this. Especially in longer texts, contradictions can occur and a narrative thread can suddenly be broken (Elkins & Chun, 2020, p. 3), whereby the model contradicts itself and produces incoherent results (Dou et al., 2022, p. 17). This makes the model an "unreliable narrator" (Dale, 2021, p. 117) with doubtful credibility. In most cases, fact-checking is necessary to verify statements (Dou et al., 2022, p. 17).

Since GPT-3 was trained using language data of human origin (Elkins & Chun, 2020, p. 13), the model also includes any errors or problematic statements it contains. Thus, it is generally possible that it adopts poor language practices and common mistakes reproduce. As a minor problem, this results in a negative influence on the linguistic quality and grammar (Dou et al., 2022, p. 17). Words could be used incorrectly, such as confusing ‘if’ and ‘whether’. Also, proverbs or analogies could be used incorrectly (Floridi & Chiriatti, 2020, p. 691). However, far worse consequences of this effect turn out to be learned biases, inbuilt prejudices and the problematic outcomes associated with them. After all, the model also contains all “of humanity’s worst tendencies” (Floridi & Chiriatti, 2020, p. 689) and the corresponding linguistic expressions (Dale, 2020, p. 116). Racist, sexist and misogynistic statements and stereotypes can thus be reproduced (Elkins & Chun, 2020, p. 3), especially if a request provokes such statements (Floridi & Chiriatti, 2020, p. 689). As (Abid et al., 2021) and Brown et al. (2020, p. 38) showed, GPT-3 associates stereotypical beliefs with certain religious communities. Lucy & Bamman (2021) showed similar effects regarding gender bias and stereotypes, whereby it tends to attribute more power to male characters as an instance. OpenAI themselves also point out such risks and social biases and recommend monitoring in this regard (OpenAI, n.d.-c). Furthermore, hate speech against any minorities or political opponents could be produced easily and quickly with the help of corresponding requests (McGuffie & Newhouse, 2020, p. 2). GPT-3 also knows of many narratives for producing content containing conspiracy beliefs or fake news (Zellers et al., 2020) and can output extremist texts of all kinds (McGuffie & Newhouse, 2020, p. 2). While such content could very easily be generated in high volume (McGuffie & Newhouse, 2020, p. 1), the high text quality could help to make such misinformation seem credible (Eliot, 2021, p. 6).

Thus, evaluating the output of GPT-3 turns out to be a double-edged sword. The quality of the output ranges from trivial nonsense to excellent writing (Elkins & Chun, 2020, p. 2). On the one hand, its biggest failures give the impression that the model is talking without any clue (Elkins & Chun, 2020, p. 12). On the other hand, with a little editing and selection, extremely impressive results can be generated, including all the strengths mentioned above. The quality of texts is largely defined by how many chances the model is given to complete a prompt and which result is finally selected. The more human intervention, the greater are the chances of a good output (Dale, 2021, p. 116).

The discussion reveals that, even though GPT-3 can impress with some very convincing results, it is neither actually intelligent, clever, self-aware, nor particularly empathetic, sensitive (Floridi & Chiriatti, 2020, p. 690) or moral. Due to the error-proneness and ethical concerns, the relevance of

human monitoring and evaluation of the output needs to be emphasised. While the impressive examples are selected and published, all the failed attempts, nonsensical results and potential corrections made have little public visibility (Dale, 2021, p. 116; Elkins & Chun, 2020, p. 8). Yet it is precisely this human control of the output that is essential for the further use of the text or its publication.

## *II State of research*

### 4 Readers' perception of computer-generated text

After having discussed the general abilities and shortcomings of AI-based NLG systems, the state of research on readers' perception of machine-produced text will be depicted. Adapted from the approach in the study by Wu (2019) on the perception of automated news stories, the theory of cognitive authority according to Wilson (1983) serves as a basic conceptual framework. It suggests that the perception of journalistic content and the trust that recipients place in it depend on several factors (Wilson, 1983; Wu, 2019, p. 6). Among others, the quality perception and evaluation of texts depends on the mere quality of the content, but also on the "personal authority" (Wu, 2019, p. 6) of the declared author of a text. The pure text quality thus refers to the influence of the actual text source and the question to what extent actual computer-produced texts are perceived differently compared to human-written texts. The perceived authority of the stated author refers to the influence of the declared source, in other words to the question of what influence the disclosure of the machine origin of a text has on the perception of its quality.

A detailed presentation of previous research on the influence of the actual source and the declared source follows in the next two chapters. First, findings on the general reader perception of computer-produced texts compared to human-written texts are presented (influence of the actual source). In this regard, it becomes evident, that simple fact-based texts produced by template-based software (see Chapter 2.1) are perceived as less readable than corresponding human texts. In terms of perceived credibility and journalistic expertise, they perform more or less equally to human versions (Graefe et al., 2018; Graefe & Bohlken, 2020; Jia, 2020; Wu, 2019). Articles and poems written by advanced AI-based language models (see Chapter 2.2) can no longer be reliably distinguished from human versions by readers (Brown et al., 2020; Clark et al., 2021; Köbis & Mossink, 2021), which suggests a high text quality. Second, the influence of the declared source of the texts on the perception and evaluation is discussed. In this respect, it can be summarised that, although not significant in all cases, most studies to date show a negative effect on quality perception when articles are declared to be computer-written. Skepticisms towards the quality and trustworthiness of allegedly machine-made or automated texts can be noticed.

## 4.1 Influence of the actual text source

In the following, the current state of research on the general perception of computer-written text is described. The chapter deals with the influence of the actual text source (human or computer) on the text quality perception, without taking into account the potential influence the declared source. First, findings on the perception of text outputs of template-based systems will be depicted, followed by the perception of the output of more advanced, AI-based language models. This distinction is reasonable because the two technical implementations of automating writing tasks differ significantly which is why the output quality and its perception are only comparable to a limited extent (see Chapter 2.1).

Up to date, several studies examined the perception of computer-written texts in different languages. The first ones were simple, template-based language systems that used different, predefined sentence templates, filled corresponding gaps with appropriate data and added some meaningful sentence parts where necessary (Dale, 2020, p. 484). In a first pilot study on the topic, Clerwall (2014) investigated readers' perceptions of template-based articles in English among 46 Swedish students (Clerwall, 2014, p. 524). Computer-written news texts turned out to be slightly more accurate, trustworthy, objective and informative, but were also perceived as more boring. At the same time, human texts were perceived as "more coherent, well written, clear, less boring, and more pleasant to read" (Clerwall, 2014, p. 525). However, these differences in readability in favour of the human version were only significant for the item "pleasant to read" (Clerwall, 2014, pp. 525-526). The study thus already showed that journalistic articles generated by a computer are barely perceived differently from human works and that the differences are small (Clerwall, 2014, p. 527).

In the course of the increasing use of automated text generation software, several studies on the perception of computer-generated news followed in different countries. In a meta-study, Graefe & Bohlken (2020) summarised the core findings from twelve experimental studies published between 2017 and 2020 (Graefe & Bohlken, 2020, pp. 51-52). The aggregated test data from experiments conducted primarily in English, but also in German, Chinese, Finnish and Korean included a total of 4473 subjects (Graefe & Bohlken, 2020, pp. 53-54). All studies examined the perception with regard to the evaluation criteria 'credibility', general 'expertise' (quality) and 'readability' (Graefe & Bohlken, 2020, p. 51). Regarding credibility, there was no difference in how human-written and computer-written news articles were perceived. The quality (expertise) of human-written news was rated only slightly better than automated articles. The readability of the human-written texts, on the other hand, was rated significantly better compared to the computer-written ones, with a big

difference (Graefe & Bohlken, 2020, p. 55). In the following, the most relevant studies included in the meta-study by Graefe et al. (2018), Wu (2019) and Jia (2020) will be discussed in more detail.

The findings of the meta-study by Graefe & Bohlken (2020) are confirmed in the so far most extensive experimental online study on the perception of computer-generated news articles, conducted by Graefe et al. (2018). 986 German-speaking participants were consecutively assigned two articles with randomised actual sources, that is, either written by a human or by an algorithm. In addition, one article was randomly displayed as written by a human author and one as written by a computer in the author line to measure the influence of the claimed source (Graefe et al., 2018, p. 604). Participants then rated the articles regarding readability, credibility and journalistic expertise. Regardless of the alleged source, articles written by a computer software were judged to be slightly more credible and were said to have slightly higher journalistic expertise. However, these differences turned out to be very small and not significant. In terms of readability, readers preferred the human texts and rated them significantly better than the computer-written texts. In cases where the source of the articles was correctly indicated, the difference in favour of the human articles was particularly high (Graefe et al., 2018, p. 604).

Regarding credibility, Wu (2019) also found that automatically produced news stories in English were rated as more credible and objective than the human version (Wu, 2019, p. 1). In an online experimental setup, the perception of template-based computer-written texts was investigated (Wu, 2019, p. 9). 370 participants from the US were presented one human and one computer-written article in either a message only setting without declared authors, or in a setting that explicitly stated the real source (Wu, 2019, pp. 8-10). The computer-written news stories were significantly rated as more credible and less biased in both conditions. These findings imply an increased perceived objectivity compared to human texts, which fits with the general assumption that algorithms can function more objectively than humans and accordingly show significantly less bias (Wu, 2019, p. 12).

By contrast, another online study by Jia (2020) on the perception of automatically written journalistic articles in Chinese, could not show any significant difference in the credibility of human versus computer-written texts, similar to Graefe et al. (2018). 125 Chinese participants successively read either four human or four AI texts, each on the same topic, without being informed about the origin of the texts (Jia, 2020, p. 2616). They then evaluated the credibility, readability and expertise of the articles using various items (Jia, 2020, p. 2617). Significant differences were only found for perceived expertise. While human articles in the areas of civic & finance news had a higher

perceived expertise, the perceived expertise of automated news was higher for conference news (Jia, 2020, p. 2619). The result is therefore ambiguous. In a follow up experiment, the same study was conducted again, only this time each of the 308 participants rated both the four human and the four computer-written texts (Jia, 2020, p. 2620). The human articles were rated significantly better than the computer-written articles in the areas of readability and expertise. In neither version of the study was a significant difference found in the credibility of human-written versus computer-written texts (Jia, 2020, p. 2623).

The results demonstrate that computers are generally capable of very convincingly producing simple journalistic texts in various languages, whereby the perceived quality is hardly any different from human output. The current areas of application, which were also tested in the studies, mainly include relatively simple news articles with a pure reproduction of facts. It comes as no surprise that the performance of computer-written content is often on a par with that of human authors, especially in terms of credibility and expertise. Template-based systems “strictly follow standard conventions of news writing” (Graefe et al., 2018, p. 604), whereby mainly facts are strung together and linguistically rather simple and prefabricated formulations are used (Graefe et al., 2018, p. 604). More elaborate plots, complex narratives and reasonable argumentation involving logic relations are only possible with more advanced systems (see Chapter 2.2). State-of-the-art language models, such as the GPT-3 model studied in this thesis, are based on self-learning mechanisms and do not require pre-written templates (Tewari et al., 2021, p. 11).

Concerning such advanced AI systems, Gunser et al. (2022) investigated whether German-speaking readers can distinguish poems and literary short texts completed by an AI from the corresponding human originals. For this purpose, on the basis of 18 original poems and short texts, for example by Kafka or Hölderlin, one completion was generated by a professional human author and one by an AI. GPT-2, the predecessor of the language model GPT-3, was used for this purpose. In a first study, 120 participants were shown nine AI-written and nine human completions of the texts. A second study with 302 participants was conducted in exactly the same way, with the only difference that this time the human texts used were not completions, but the poets’ original works. For each text, the readers indicated whether they thought it was AI-written or human and how sure they were of their decision (Gunser et al., 2022, pp. 1746-1747). In addition, the stylistic text quality was surveyed by asking how “well-written, inspiring, fascinating, interesting, and aesthetic” (Gunser et al., 2022, p. 1746) they perceived the works. Overall, the AI versions were perceived as worse than both the original and the human completion in terms of all five evaluation categories (Gunser et al., 2022, pp. 1747-1748). These findings are consistent with the finding for template-based texts

that human texts have better readability (Clerwall, 2014; Graefe et al., 2018; Jia, 2020). Readers still prefer human texts for readability and pure language quality. Poems are also the significantly more linguistically challenging text genre compared to relatively simple news articles, though. Despite the perceived quality differences in all evaluation criteria, however, readers found it difficult to correctly distinguish the AI versions from the two human versions (Gunser et al., 2022, p. 1749). In both studies, almost 60% of the AI-written poems were correctly recognised as such. Although this is a slightly better result than random guessing, it does not show a reliable identification of the source (Gunser et al., 2022, pp. 1746-1747). However, on average, the subjects believed that they had correctly identified the texts. Subjects were thus often overconfident in their classification of the texts (Gunser et al., 2022, p. 1748).

A similar study on the perception of GPT-2-written poems in English was conducted by Köbis & Mossink (2021). The aim was to find out whether people can distinguish between AI-written and human-written poems and which version they prefer. For this purpose, similar to the procedure in Gunser et al. (2022), beginnings of human original poems were given as input, from which GPT-2 generated several completions. The first part of the study compared the perception of poems by untrained human authors with the perception of poems by a medium-sized GPT-2 model. The final AI poems used were selected by humans. 192 participants were then shown 10 pairs of poems, one from an AI and one from a human. Participants were also randomly divided into two groups (opacity & transparency). One group only knew that one of the poems was AI-written and one was human-written (opacity), and the other group was told which of the two was which (transparency). Both groups then indicated which of the poem versions they preferred. In addition, the opacity group was asked to identify the AI-written version and to indicate on a scale of 1 to 100 how confident they were in their decision. The second part of the study was conducted under the same conditions with a total of 384 participants, with the only difference that poems from professional human authors were used and the complete, most powerful model of GPT-2 was used to generate the AI versions. The pool of AI poems finally used was selected once by humans and once at random, and it was randomly varied which participant received which pool (Köbis & Mossink, 2021, pp. 3-4).

In both parts of the study, the human-written was preferred over the GPT-2-written poetry overall. This preference for the human-written texts was even more evident in the second part of the study when the AI versions were selected at random rather than by humans (Köbis & Mossink, 2021, p. 8). When detecting poems written by computers, participants were unable to distinguish the AI from the human-written versions better than at a 50% chance level (Köbis & Mossink, 2021, p. 5), even compared to renowned human authors (Köbis & Mossink, 2021, p. 10). The second part of



the study showed that only the poems that had been pre-curated by humans were not reliably recognised as such. In contrast, the experimental group with randomly selected poems succeeded in identifying the AI version more often than average (Köbis & Mossink, 2021, p. 9). Furthermore, in analogy to the results of Gunser et al. (2022), there was some evidence of self-overestimation in the recognition of computer-written texts. Most readers rated their own ability to recognise AI texts higher than was actually the case (Köbis & Mossink, 2021, p. 6).

The results show that the AI can write appealing poems that can hardly be distinguished from the results of professional poets. Human participation in the selection process of the texts turned out to be extremely important (Köbis & Mossink, 2021, p. 11). However, human-written poems are still preferred overall according to both the studies of Gunser et al. (2022) and Köbis & Mossink (2021). This fits with previous findings about template-based NLG news articles, which were perceived as less linguistically proficient and readable (Clerwall, 2014; Graefe et al., 2018; Jia, 2020).

To date, three studies involving AI-written journalistic content exist. Tewari et al. (2021) conducted an experiment on the perception of GPT-2 articles. GPT-2 was given the first sentences of a human article as an input, which already indicated a certain political stance for the article to be generated. Like this, several AI-written articles of about 400 words length were produced on the topics of climate change, vaccinations and politics. Articles with conservative and liberal leanings were generated for each topic. The human-written versions on the same topics were taken from American news outlets, while some also featured conservative and some showed liberal leanings (Tewari et al., 2021, p. 4).

In a laboratory experiment, 30 participants were randomly presented with a total of twelve articles without indicating the actual source, half of which were written by GPT-2 and half by a human. For each article, the participants indicated how credible they found it and whether they believed the article was written by a computer or a human. Results show that the AI-written texts were rated as significantly less credible compared to human texts (Tewari et al., 2021, p. 7). The AI texts were correctly identified as such in about half of all cases, while human articles were correctly identified in about 70% (Tewari et al., 2021, pp. 6-7). Readers thus seemed fairly capable of distinguishing between human and machine texts and revealed a rather critical attitude towards computer-produced texts overall (Tewari et al., 2021, p. 11). The low rating of credibility in contrast to the template-based articles discussed at the beginning makes sense because of the AI language model tested here, which is not necessarily fact-based (see Chapter 3.3). In addition, pulse and skin conductance were recorded to determine emotional tension during reading (Tewari et al., 2021, pp. 5-6). The peaks in these data were equally distributed during the reading of machine and human-

written texts with no significant differences, showing that readers have the same emotional involvement while reading for both article versions (Tewari et al., 2021, p. 7). However, due to the small sample size, these results are less robust than other studies.

The second study on the question of whether readers can distinguish AI-written articles from human ones was conducted by Clark et al. (2021) with GPT-2 and GPT-3 texts. For this purpose, a total of 780 participants from the USA indicated on a scale of one to four how human-like they considered a total of five presented text passages that were written either by the AI or by a human (Clark et al., 2021, p. 2). In three different experimental groups, texts from the genres stories, news articles and recipes were examined. Results show that readers could not reliably distinguish GPT-3 texts from human ones (Clark et al., 2021, p. 1). In 49.9 % of the cases, readers correctly guessed the source, which is equivalent to random guessing. For texts from the older and less powerful GPT-2 model, the rate of correct identification was slightly higher, at 57.9% (Clark et al., 2021, p. 4). In addition, participants were asked for reasons for their assessment. This revealed that readers underestimated the competences of the language models. The judgements were also mostly based on superficial text properties (Clark et al., 2021, p. 9). Human characteristics or intentions in the AI-written texts were often referred to as reasons. Many subjects assumed that the texts were human, because they did not believe that a computer was capable of such human language characteristics (Clark et al., 2021, p. 7).

These results are supported by a third, very similar experiment conducted by Brown et al. (2020), in the course of the development of GPT-3 by OpenAI themselves. The aim was to find out whether people can distinguish GPT-3-written news articles from human versions. Using a total of nine GPT-3 submodels of increasing ability, they generated completions for news articles in English based on 25 titles and subtitles. For each GPT-3 version, 80 US participants were presented with the 25 articles in randomised order, half of which were AI-written and half human-written (Brown et al., 2020, p. 47). They then indicated how likely they considered the text to be human or machine-written (Brown et al., 2020, p. 25).

The test was carried out using texts with a length of about 200 words. The texts produced by the least capable model could be correctly identified by the test persons in 86% of the cases. For the state-of-the-art version of GPT-3, the rate of correctly classified texts was only 52%, which is only slightly better than random guessing. As expected, the larger and the more powerful the GPT(-3) model used, the more difficulties participants had in correctly identifying the source. Although the larger the model, the more time readers took to make their guesses, the ratio of correct guesses of the source for the most capable model is near pure chance (Brown et al., 2020, p. 26). Larger

models of GPT(-3) thus write news articles that are more difficult to distinguish from human versions (Brown et al., 2020, p. 47). Since these findings only relate to articles of about 200 words in length, the same experiment was carried out again for GPT-3 news articles of 500 words. Using twelve human news articles from Reuters, AI versions were again generated with the most powerful GPT-3 model and with a weaker control model. 160 US Americans were shown twelve articles per model, with half of the texts human and half AI-generated, as in the previous experiment (Brown et al., 2020, p. 48). For the best-performing model of GPT-3, the rate of correctly identified articles was still just above chance level at 52%. Even for longer texts around 500 words, GPT-3 seems to be capable of producing AI texts that can hardly be recognised as such by humans (Brown et al., 2020, p. 26). The rate of about fifty percent of correctly identified AI texts is in line with the results of Gunser et al. (2022) and Köbis & Mossink (2021) for AI-generated poems. Even though an exact quality rating was not queried, these results indicate a high text quality of the GPT-3 news articles.

In summary, it becomes evident that texts produced by template-based software (see Chapter 2.1) are perceived as linguistically less proficient and readable than corresponding human texts (Clerwall, 2014; Graefe et al., 2018; Graefe & Bohlken, 2020; Jia, 2020). In terms of perceived credibility and journalistic expertise, on the other hand, they perform equally to slightly better in some cases (Graefe et al., 2018; Graefe & Bohlken, 2020; Jia, 2020; Wu, 2019). Articles and poems written by advanced self-learning language models (see Chapter 2.2) can no longer be reliably distinguished from human versions by readers. This is the case for both GPT-3 and its predecessor GPT-2, with a rate of correctly identified text sources (AI or human) of around 50% (Brown et al., 2020; Clark et al., 2021; Köbis & Mossink, 2021).

## 4.2 Influence of the declared text source

After having presented the current state of research on the general perception of computer-written text, the influence of the declared source on the quality perception shall be depicted and discussed. The perception of news does not only depend on the content of a given article, but also on the perceived article source Wilson (1983), that is, the “personal authority” (Wu, 2020, p. 6) of a text. Possible effects on the perception of texts when it is revealed that they were written by a computer are discussed.

Van der Kaa & Krahmer (2014) were the first to investigate the influence of the declared source on the perception of computer-written texts in an experimental design. 232 dutch news consumers

were presented with simple, template-based articles. Each participant read one computer-written texts from the field of sports or finance, which was either claimed to be human or computer-written. No significant difference was found between the different sources in terms of perceived trustworthiness and expertise (e.g. credibility) (van der Kaa & Kraemer, 2014).

The already introduced study by Graefe et al. (2018) likewise found barely any differences in the evaluation of the texts when varying the specified article source (human or computer). The experiment involving 986 test persons, revealed that stating a text was human had only a small positive effect on the quality rating. Regardless of their actual source, all articles declared as written by a human were rated better in all evaluation criteria (credibility, expertise and readability), although these differences were not all statistically significant (Graefe et al., 2018, p. 603). Only such articles written by a computer that were falsely labelled as being written by a journalist were rated significantly better in terms of readability (Graefe et al., 2018, pp. 603-604).

Waddell (2018), by contrast, provided stronger evidence for a negative influence if a computer was stated as an article source. In an experiment, the impact of stated authorship of computers on the perceived credibility of news was investigated. 129 participants from North America were given a human news article to read, in which either a so-called robot reporter or a human reporter was named as a source in the byline of the article (Waddell, 2018, p. 239). Participants then rated the texts regarding credibility, quality, newsworthiness and representativeness. News content, when attributed to a computer origin, turned out to be perceived as overall less credible than when attributed to a human journalist (Waddell, 2018, p. 248). Moreover, allegedly computer-written news were rated poorer in quality and perceived as less newsworthy and representative (Waddell, 2018, p. 241). As the study has a relatively small sample, these findings are only comparable to a limited extent with those of van der Kaa & Kraemer (2014) and Graefe et al. (2016) (Waddell, 2018, p. 250).

However, the previously mentioned study of Wu (2019) with 370 participants discovered similar negative effects when researching on the influence of the declared source in an online experimental setup with computer-written texts (Wu, 2019, pp. 8-9). It turned out that the declaration of source and author has a negative influence on the perception of computer-written articles. Computer-written articles were rated as more objective and credible than human written news stories in a message only condition, without a declared source (Wu, 2019, pp. 11-12). The evaluators thus showed more skepticisms towards the articles once they knew that they were computer-written. There was no significant difference for the evaluation criterion of 'bias' (Wu, 2019, p. 1). This result

is consistent with the findings of Graefe et al. (2018) suggesting that human source attribution can be associated with a somewhat better evaluation of the texts (Wu, 2019, p. 14) as well as the findings of Waddell (2018).

A certain amount of skepticism towards software that produces texts and its capabilities was also shown in the study by Jia (2020) for the Chinese region. In the study described above (see Chapter 4.1), the expectation of the article quality prior to reading was also surveyed. Both the expectations of readability and expertise turned out to be significantly higher for the human-written texts than for the computer-written texts. The expectations on credibility were slightly lower for human texts than for computer-written texts although without significant differences (Jia, 2020, p. 2621). A comparison with the subsequent actual evaluation of the texts showed that human articles were overestimated in the pre-expectation with regard to readability and expertise, but credibility was underestimated. Automated articles tended to be generally underestimated in the pre-expectation, whereby significant differences could only be demonstrated for credibility and expertise (Jia, 2020, p. 2624).

The previously described meta-study by Graefe & Bohlken (2020), which summarised twelve experiments on the perception of template-based, computer-produced texts also included results on the influence of the declared source on article perception. In line with the four studies above, they found that articles claimed to be written by a human author performed better than those claimed to be written by a computer on average across all combined study data. Readers overall rated credibility, quality and readability higher “if they *thought* that they read a human-written article” (Graefe & Bohlken, 2020, p. 57). This effect was medium-sized in the case of credibility, small in the case of quality and very large in the case of readability. The personal authority of the declared source thus has a major effect on perception (Graefe & Bohlken, 2020, pp. 56-57), whereas the revelation of computer-authorship has negative effects.

A similar pattern of skepticism towards computer-generated content can be found for AI-generated texts. As described above, the experiment of Tewari et al. (2021) on the perception of GPT-2-produced texts generally revealed a certain level of skepticism towards AI articles. This was also reflected in poorer text ratings if readers assumed they were reading an AI-written article (Tewari et al., 2021, p. 7). Although no source was indicated for the articles in this study, readers rated texts they thought were written by machines as significantly less credible, regardless of the actual source. Especially for articles with a political leaning, the recipients apparently did not trust that computers can reliably write them.

The introduced research on AI-written poetry by Köbis & Mossink (2021) generally found that the test persons showed some skepticism towards the actually computer-written content and overall preferred human-written poems (see Chapter 4.1). For the experiment, a total of 576 participants were shown 10 pairs of poems, one by an AI and one by a human. Thereby they were randomly divided into two groups (opacity & transparency). One group only knew that one of the poems was AI-written and one was human-written (opacity), and the other group was told which of the two was which (transparency). Both groups then indicated which of the poem versions they preferred. Participants' preference for human-written poems did not increase when they knew about the actual origin of the texts (Köbis & Mossink, 2021, p. 5). The source declaration thus had neither a positive nor a negative influence on the perception of the AI poems.

Furthermore, in a recent experimental study, Longoni et al. (2022) investigated how the declared source of messages influences the perceived news accuracy. In a sample of news consumers representative of the US population, a general aversion towards AI-generated news was found. In a first online experiment, 3029 participants were randomly presented with either some news items declared as written by an AI or some news items declared as written by a human journalist (between subjects). The news items consisted of a real headline and a suitable image, which were both taken from actual news outlets. A second experiment with 1005 participants was conducted in a similar manner, with the exception that participants each read a set of news articles that contained both of the possible indicated sources (within subjects) (Longoni et al., 2022, p. 98). In both experimental setups, readers believed the content significantly less in the cases in which the AI authorship was revealed. This effect was due to generally lower trust in the output of an AI. In addition, participants were asked about personal background and characteristics related to demographics, religion, political views, information habits and risk assessment (Longoni et al., 2022, p. 99). The effect was evident for all different personal characteristics (Longoni et al., 2022, p. 101). Revealing that an AI had written a news item thus had a negative effect on perceived credibility and news accuracy.

In conclusion, the findings on the influence of the declared source are not entirely unambiguous to date. Nevertheless, negative effects on the perceived text quality can be observed on average if no human author is indicated as the source. Although not significant in all cases, most studies to date show a clear negative effect on quality perception when articles are claimed to be computer-written (Graefe & Bohlken, 2020; Longoni et al., 2022; Tewari et al., 2021; Waddell, 2018; Wu, 2019). Skepticism towards the quality and trustworthiness of allegedly AI-written or automated texts can be noticed. This is consistent with the finding of Clark et al. (2021) that readers tend to underestimate the competences of language models (see Chapter 4.1).

### *III Empirical study*

## 5 Research gap and open questions

The state of research presented makes evident that computer-written content can generally compete with the products of human authors in terms of quality perception by media consumers. However, some research gaps become apparent.

First, the capabilities of language models will continue to improve as technology advances. In this process, they will come ever closer to the capabilities of humans in all areas of application and reader perception will continue to change as a result (Graefe et al., 2018, p. 605). Even for complex use cases readers will find it increasingly difficult to distinguish computer-written texts from human works in terms of quality. In particular, outputs from AI applications such as GPT-3, which can be freely used for a wide variety of language purposes (OpenAI, 2020), hold great potential. Therefore, the perception of AI-written texts should be further investigated, which is the basic motivation of the empirical part of the paper.

Second, many of the previous studies on the topic examined the perception of simple, template-based reports, which mainly reflect objective facts and clear information. These texts do not contain any free expression of subjective arguments, for example on political issues, and thus no controversies or debatable positions. However, with the emergence of self-learning language systems like GPT-3, more elaborate comments on various topics are possible. Studies on GPT-2 and GPT-3 output carried out so far concentrated on literary texts and poems (Gunser et al., 2022; Köbis & Mossink, 2021). Brown et al. (2020) and Clark et al. (2021) did investigate whether readers could still distinguish simple news reports written by GPT-3 from human versions, although a detailed quality assessment based on concrete criteria was not carried out. The perception of opinion statements by an AI, such as in a journalistic commentary, is therefore hardly researched (Tewari et al., 2021, p. 3). Only Tewari et al. (2021) conducted a small laboratory experiment in this regard based on GPT-2-generated articles (see Chapter 4.1). Furthermore, it remains to be seen how the much more advanced successor GPT-3 would perform (Tewari et al., 2021, p. 12).

Third, it became apparent that although computer-written texts have already been studied in a wide variety of languages, German has so far played a minor role. The only comprehensive study of automated texts in German to date was conducted by Graefe et al. (2018). For advanced AI applications such as GPT-3, studies have mainly been conducted on output in English (Brown et al., 2020; Clark et al., 2021; Gunser et al., 2022; Köbis & Mossink, 2021; Tewari et al., 2021). Especially considering the much smaller training data set of GPT-3 in German compared to

English (see Chapter 3.1), an investigation of the quality perception of German language output is of interest. It is still open how the languages less represented in the language model perform.

Fourth, previous research showed that readers sometimes perceive computer-written texts more critically and evaluate them more poorly with regard to some categories, as long as they know about their machine origin. Some recipients thus revealed a certain degree of skepticisms (Graefe & Bohlken, 2020; Jia, 2020; Tewari et al., 2021; Waddell, 2018; Wu, 2019) towards computer-written texts and underestimated their competences (Clark et al., 2021). The level of trust or skepticism towards machines or AI is closely connected to the general acceptance of the corresponding technologies (Tschopp, 2022, p. 331). It can be assumed that the indication of the author being an ‘artificial intelligence’ provokes different associations and attitudes towards the machine and its text, similar to the case with robots (Wullenkord, 2017). Yet, the role of the personal attitudes towards computer authors or AI when perceiving computer-produced content has not yet been examined. The aspect of the reader’s preconceptions will therefore be further explored in this paper. Therefore, the personal attitudes towards AI of the participants will be surveyed in order to investigate whether and how positive or negative associations towards AI are related to positive or negative ratings of machine-authored content (see RQ3).

One of state-of-the-art language models, GPT-3, has not yet been tested for opinion pieces in journalism among the German-speaking population. The following research questions and related hypotheses follow from the research gaps outlined above.

The general objective of this thesis is to investigate how readers perceive and evaluate AI-written articles, which is subject of **RQ1: ‘What influence does the actual source of the articles (AI or human) have on the perception of journalistic text quality and trust in the text?’**

Considering the previous research results, it is to be expected that AI articles perform somewhat worse than the human versions. It turned out that journalistic articles created by template-based language software were perceived more negatively overall by readers than human texts. Furthermore, journalistic texts with political leanings written by the AI GPT-2 in the study by Tewari et al. (2021) turned out to be significantly less credible than human versions. The research on GPT-2-written literature and poems in German by Köbis & Mossink (2021) and Gunser et al. (2022) also showed that readers rate the readability and language quality of human works higher. For English GPT-3 news articles, Brown et al. (2020) and Clark et al. (2021) showed that readers could not distinguish reliably between human and machine written articles. Although this indicates a human-like quality of the GPT-3-written articles, a proper quality assessment in comparison to



human texts by the readers was not conducted. Moreover, while the articles examined were simple journalistic news items that did not contain any explicit personal view, the texts tested in this paper should contain opinion statements and arguments. From a technical point of view, the amount of training data used for training GPT-3 is also much smaller in German than in English which theoretically makes for poorer performance. Both factors make the task at hand more complicated for GPT-3 than in the case of Brown et al. (2020) and Clark et al. (2021). It is therefore assumed that GPT-3 texts perform worse than the human versions in terms of perceived trust and journalistic text quality, which leads to the following hypotheses.

H1.1: Actual AI-written articles are evaluated more negatively than the actual human versions in terms of overall journalistic text quality.

H1.2: Actual AI-written articles are perceived more negatively than the actual human versions in terms of trust in the content.

Moreover, the role of the declared source is to be investigated, which leads to **RQ2: ‘What influence does the declared source of the articles (AI or human) have on the perception of journalistic text quality and trust in the text?’**

Regarding this question, the following hypotheses are assumed.

H2.1: Texts declared as AI-written are evaluated more negatively than the human-declared text in terms of overall journalistic text quality.

H2.2: Texts declared as AI-written are evaluated more negatively than the human-declared text in terms of trust in the content.

These assumptions are reasonable because of the aforementioned skepticism of readers regarding computer-generated text, which was revealed in various studies (see Chapter 4.2). A tendency towards more negative ratings is noticeable. This applies for template-based NLG, as well as AI-solutions (Graefe & Bohlken, 2020; Longoni et al., 2022; Tewari et al., 2021; Waddell, 2018). Although readers’ perceptions of computer-written content and public awareness could change rapidly (Graefe & Bohlken, 2020, p. 58), given the currency of most of the findings, it seems unlikely that this has already happened in the short time since the studies were conducted. It is therefore to be assumed that a certain skepticism towards computer-written content still prevails

and that informing readers about the fact that articles were written by an AI has a negative impact on the perception of general journalistic text quality and trust.

Proceeding from the question of the influence of the stated source, it is to be examined with **RQ3: ‘What role does the personal attitude towards AI play when perceiving and evaluating AI-generated texts compared to human written texts?’**

Two points are hypothesized here.

H3.1: Attitudes towards AI show a positive correlation with the perception of articles declared as AI-written.

It can be assumed that the reason for a negative evaluation of supposedly AI-written texts is partly due to negative personal attitudes or associations of the recipients towards AI and resulting skepticism. Accordingly, a positive correlation of the personal attitudes towards AI with the mean evaluations of allegedly AI-written articles is hypothesized. In other words, overall mistrustful attitudes towards AI go hand in hand with poorer ratings for allegedly AI-written articles, while more positive attitudes are related to more positive ratings of such articles.

H3.2: Attitudes towards AI show no correlation with the perception of articles declared as human-written.

Conversely, it cannot be assumed that there is a correlation between attitudes towards AI and the ratings of texts that are supposedly written by humans. Neither a positive nor a negative correlation between attitudes and the ratings of allegedly human-written texts would make sense. It will therefore be assumed that attitudes towards AI are unrelated to the ratings of allegedly human texts. Consequently, this hypothesis also verifies whether a possible correlation actually only exists for article versions allegedly written by AI.

## 6 Methodology

After having outlined the research gaps, the following chapter will explain the methodological approach used to answer the questions raised. The study investigates how AI-generated articles are perceived compared to corresponding human originals. Therefore, a possible causal influence of the variation of the real text source (AI or human) on the quality perception of articles needs to be tested. Given this background, an experimental research design is suitable for the project. If implemented correctly, it is best suited for testing hypotheses about causal relationships (Diekmann, 2005, p. 290; Döring & Bortz, 2016, p. 194). In the following, the experiment as a method will therefore first be described in general, in order to subsequently explain the concrete experimental implementation for the present research objective. In addition, the operationalisation and measurement of the article perception and evaluation as well as of the personal attitudes towards AI will be depicted. Afterwards, the procedure for generating the AI-written articles to be tested (experimental treatments) are then explained. Ultimately, the revision of the survey in the course of a pre-test, the sample selection and the procedure for the analysis of the collected data is discussed.

### 6.1 Experimental trials

Experimental studies are generally characterized by the fact that they allow conclusions about actual causalities (Bryman, 2012, p. 58). While other methods such as questionnaires can only identify general relationships between variables (Gnambs & Strassnig, 2007, p. 232), an experiment is able to determine clear causality. Thus, the direction of effect from one variable onto another can be determined (Diekmann, 2005, p. 303). The exact relationship between cause and effect can be determined, in other words, which causes lead to which observable phenomena (Döring & Bortz, 2016, p. 193). A hypothesized effect is accordingly measured as a dependent variable, with potentially causal factors being measured as an independent variable. The objective of an experiment is thus to investigate possible causal effects of an independent variable on a dependent variable (Döring & Bortz, 2016, p. 194).

In order to unambiguously test hypotheses about causal relationships, the necessary environment, including the specific relationships of cause and effect, to be tested must be artificially generated as an experimental setup (Döring & Bortz, 2016, p. 194). For this purpose, the potential cause of a suspected effect is artificially created as an experimental treatment (X). The possible effect of the treatment is measured by observation (O). In a standard experimental design, each participant is

assigned to one of two or more experimental groups. The experimental groups are each exposed to a systematically different treatment (X) as an independent variable, from which a certain effect is expected. Each subject is therefore exposed to only one experimental condition, resulting in a between-subjects design (see Figure 1) (Diemann, 2005, p. 296). Possible effects of the respective treatment on the dependent variable are then determined by measurement or observation (O) in the experimental groups (Döring & Bortz, 2016, p. 193). The experimental groups can then be compared with regard to the possibly different effects of the treatments presented in each case (Diekmann, 2005, p. 297).

**Table 1**

*Experimental set-up in a between-subjects-design*

Randomised group assignment	Independent variable	Dependent variable	
R	X1	O1	Exp. group 1
R	X2	O2	Exp. group 2

*Note.* X = manipulated treatment; O = observation (measurement). Illustration following Diekmann (2005).

In order to draw any conclusions at all about the causal direction, the introduced stimulus must chronologically precede the potential effect to be examined (Diekmann, 2005, p. 303). This preceding in time constitutes the basis for statements about causality (Sedlmeier & Renkewitz, 2018, p. 134). Accordingly, the independent variable in Table 1 causes an effect on the dependent variable, and not the other way round. Without this controlled sequence, only a general correlation between two variables could be investigated.

Furthermore, a definite attribution of an effect to a causal factor (independent variable) can be prevented by subject- or study-related confounding variables. Subject-related confounding variables include all differences in the personal characteristics of the test subjects (Döring & Bortz, 2016, p. 193), such as the degree of motivation when completing the experiment, intelligence, or the extent of prior knowledge about the respective subject area and any other possible personal characteristic (Sedlmeier & Renkewitz, 2018, p. 136). Confounding variables caused by the environment or the study design refer to possible unequal treatment of the experimental groups, for instance if groups receive different levels of intensive support or encouragement from the experimenter. But potential distractions, noise or the time of day can also be confounding variables, especially in online experiments (Gnambs & Strassnig, 2007, p. 233). Confounding variables thus represent alternative explanations for an observed effect. If, for example, many people with above-

average intelligence are assigned to an experimental group, the factor intelligence could be an alternative explanation for observed effects in addition to the independent variable to be tested. In this way, it could no longer be reliably assumed that the different treatment alone is responsible for the respective effect (Sedlmeier & Renkewitz, 2018, p. 136). Such impacts of possible confounding factors and alternative explanations must thus be eliminated and neutralized. This is ensured by establishing control or comparison groups in addition to randomly assigning the test persons to the groups (Bryman, 2012, pp. 51-52). An experimental design thus involves three conditions, which significantly determine the course of an experiment. The basic requirement is the aforementioned *group formation*, which is carried out by *randomising* the participants. In addition, the independent variables must be *manipulated* in at least one of the groups (Diekmann, 2005, p. 296). The fulfilment of these conditions is absolutely essential for the success of a true experimental design and will therefore be explained in detail.

At least two experimental comparison groups are formed (Diekmann, 2005 p. 296; Bryman, 2012, p. 58), which are treated differently (Döring & Bortz, 2016, p. 193). All subjects are initially assigned to either an ‘experimental group’ or a ‘control group’ (Bryman, 2012, p. 51), which are then exposed to different treatments, that is, versions of the dependent variable (Diekmann, 2005, p. 296; Bryman, 2012, p. 51). Differences in the observable effects of the treatments can then be compared between the experimental and control groups (Döring & Bortz, 2016, p. 194). The control group is also often referred to as a ‘comparison group’ (Bryman, 2012, p. 58) when comparing the effects of different independent variables (Bryman, 2012, p. 58). In cases where no typical control group is used, which receives a completely ineffective treatment (placebo), the term ‘comparison’ or ‘experimental group’ is more appropriate (Diekmann, 2005, p. 297). This also applies to the present study, which aims to compare the effect of two text treatments (AI-written vs. human text). It is therefore rather a matter of two experimental comparison groups.

The assignment of participants to the different experimental groups must be done on a strictly random basis (Diekmann, 2005, p. 296). Only in this way can it be assumed that the artificially formed groups are exactly comparable groups (Döring & Bortz, 2016, p. 193). If a large number of participants are assigned to experimental groups by randomisation, it can be assumed that all possible confounding variables are distributed equally over both groups on average due to chance distribution (Diekmann, 2005, p. 297). Influences of confounding variables acting in parallel to the independent variable thus occur equally in all test groups or experimental conditions, whereby no unwanted systematic interference can occur (Gnambs & Strassnig, 2007, p. 233). The potential influence of unknown variables is the same in each experimental group which makes them uncorrelated to the independent variable and neutralises the confounding effects accordingly. A

bias due to aggregation of certain confounding factors in one experimental condition is therefore not possible and any observable effect can be attributed exclusively to the prior independent variable (Diekmann, 2005, p. 297). Possible effects, that is, differences in the recorded values of the dependent variable between the two experimental groups, can therefore be attributed exclusively to the manipulation of the treatment (Bryman, 2012, p. 51).

To be able to determine differences between the groups in this manner, they obviously must be *treated systematically* different (Döring & Bortz, 2016, p. 193). This is achieved by manipulating the independent variable in the form of the treatment (Diekmann, 2005, p. 296; Döring & Bortz, 2016, p. 193). One or more treatments are shown as an independent variable, whereas the treatment received by the experimental group is systematically manipulated as compared to the ones of the comparison or control group (Diekmann, 2005, p. 296). Thus, at least two different versions of variables are used as independent variables, a manipulated and an unmanipulated one (Sedlmeier & Renkewitz, 2018, p. 134). This allows the potential impact of the manipulated variable on the subjects to be captured (Bryman, 2012, p. 50). The effects become measurable in the dependent variable (Sedlmeier & Renkewitz, 2018, p. 134). Based on the measured differences between the groups, the impact of the different independent variables can be determined (Bryman, 2012, p. 50).

**Table 2**

*Experimental set-up in a within-subjects-design*

Random group assignment	Independent variable A	Dependent variable	Independent variable B	Dependent variable
R	X1	O1	X2	O2
R	X2	O2	X1	O1

*Note.* X = manipulated treatment; O = observation (measurement). Illustration following Diekmann (2005).

As opposed to the between-subjects design, it is also possible to set up an experiment in a within-subjects design. In the within-subjects design, all subjects pass through all or several possible treatments. Again, the order of treatment and subsequent measurement of effects is essential in order to be able to make statements about causal relationships and directions of effect. Since all subjects go through all conditions, a sufficient amount of data per treatment can be achieved more effectively than in the between-subjects design. While in the between-subjects design one subject would be assigned to independent variable A and another to independent variable B, in the within-subjects design a single subject passes through both independent variables A and B (see Table 2). A between-subjects design therefore needs twice as much participants for the same amount of test data per condition (Sedlmeier & Renkewitz, 2018, p. 157).

## 6.2 Implementation as multi-factorial within-subjects online experiment

Having explained the basic design and the inherent logic of experiments, the following section deals with the concrete experimental design for testing the hypotheses given here. For the practical implementation of the experiment, an online experiment in a multi-factorial within-subjects design was used. The design was implemented according to the model of Graefe et al. (2018) and was suitable due to the effective data collection for all conditions and the randomisation options that allow for optimal control of confounding variables. In the following, it is explained as to why and how the implementation was carried out, for one, as an online experiment and, for the other, in a multi-factorial within-subjects design. The entire procedure of the experiment, including all questionnaire pages, can be found in Appendix 4.

### 6.2.1 Online experiment

The experimental setup was carried out as an online experiment. The entire experimental procedure took place online and the test subjects were presented with all the questionnaires and appropriately manipulated stimulus material using the web. All instructions, the randomised assignment, as well as the questioning of the items via questionnaire were automated and pre-programmed. A study leader as in classical laboratory experiments was thus not necessary (Gnambs & Strassnig, 2007, p. 234). The access to the online experiment was sent to potential test persons via a link, which can then be opened via their personal devices with the help of a web browser. The entire online experiment was implemented using the online survey tool Unipark ([unipark.com](http://unipark.com)).

Conducting the experiment as an online experiment suited this work, as inviting the test persons to an on-site laboratory would not have been feasible in the time available given the necessary sample size. Also, the comparatively short data collection period for online experiments helped. Most of the test persons participate in the study in the first few days after being invited. It was therefore possible to achieve a relatively extensive sample in short time (Gnambs & Strassnig, 2007, p. 237). In addition, conducting the survey on an online platform offers a high degree of automation. Data storage could be automated, as well as the randomised assignment to the experimental conditions and a varying experimental procedure (Gnambs & Strassnig, 2007, p. 236). Furthermore, online experiments typically features increased external validity, which has a positive effect on the generalisability of the results. On the one hand, this is due to the fact that online experiments can theoretically reach people from all over the world and in contrast to laboratory experiments, are not tied to a specific location. Thus, more heterogeneous samples can be reached

with less effort (Gnambs & Strassnig, 2007, pp. 236-237). On the other hand, the subjects go through the experiment in very different, personal situations and not in a strange and unfamiliar laboratory environment, which can cause bias. Due to the more familiar environment in which the data is collected, it can be considered less biased than data collected in an artificial laboratory setting when it comes to external validity. Test subjects take part in the experiment at different locations and at different times of day, which comes as close as possible to collecting data under 'real-life' conditions (Gnambs & Strassnig, 2007, p. 237). Potentially observed effects are therefore not due to the unfamiliar laboratory environment (Reips, 2002, p. 247). As mentioned above, there is no need for interaction with a supervisor, as all instructions are always given in the same, written form. Unlike for laboratory experiments, distorting supervisor effects are therefore not an issue and unequal treatment of the test persons cannot occur (Gnambs & Strassnig, 2007, p. 238). Because of these reasons, the results are usually better transferable to real situations (Gnambs & Strassnig, 2007, p. 237). The present study benefits from this because the environment in which the test persons read the text examples is relatively close to the real-life setting of the consumption of journalistic content (in private surroundings, with one's own device).

In turn, this also implies a reduced internal validity of online experiments compared to laboratory experiments. Ambient and situational confounding variables cannot be completely controlled. In contrast to the fully controllable laboratory experiment, distractions can occur due to other people, phone calls, messages or any other activity carried out in parallel. Technical equipment or potential problems may also vary among participants (Gnambs & Strassnig, 2007, pp. 237-238). This can affect the quality of the data and lead to high dropout rates. Moreover, in an unsupervised setting, the hurdle of stopping the experiment midway is much lower than in a laboratory setting. Due to the danger of a systematic dropout, this can also reduce the data quality of the results and validity. If dropouts occur systematically more frequently in certain experimental conditions than in others, this may cause strong biases. For instance, the data collected for the corresponding condition would then consist of a large number of people who are generally particularly motivated. At the same time, the data for the remaining conditions would contain an average or low level of motivation (Gnambs & Strassnig, 2007, p. 238). Motivation would turn out to be an uncontrolled confounding variable (Reips, 2021, p. 208). In order to avoid such a systematic dropout in the present experiment from the start, it was ensured that the texts were approximately the same length. The complexity of the texts is also roughly comparable per topic (1 & 2). This should prevent the incentive to drop out more often in one condition. Additionally, a warm-up was applied. Before the experimental setup, the participants were first asked opening questions on their prior knowledge on the topic and their personal attitudes towards AI (Gnambs & Strassnig, 2007, p. 240). People who have no



serious intention of participating in the experiment may then drop out already in the beginning, before entering the actual experiment. The gathered data was subsequently also checked for frequent dropouts in individual conditions (Gnambs & Strassnig, 2007, p. 239) (see Chapter 6.8). Another problem can be intentionally given wrong answers, such as by people who just randomly click through all the pages. Technical problems can theoretically occur due to different browsers and operating systems (Gnambs & Strassnig, 2007, p. 238), but they are negligible because the experiment was implemented with the universally usable online software Unipark, which works in every browser.

### 6.2.2 Multi-factorial within-subjects design

The online experiment was set up as a multi-factorial 2x2x2 within-subjects design. Accordingly, the independent variables (factors) topic, actual source and declared source were each varied in two stages (Sedlmeier & Renkewitz, 2018, p. 170). The study was based on journalistic opinion pieces on two different topics (1 & 2). For both topics, a version written by a human and a version written by an AI were provided (real source), each with a correctly and an incorrectly stated source (declared source). This resulted in eight possible treatments or article versions, as depicted in Table 3.

**Table 3**  
*Possible article versions as experimental treatments*

Real source	Declared source	Topic	Treatment
<b>Human-written</b> article	correctly declared as <b>human-written</b>	A	1
		B	2
	wrongly declared as <b>AI-written</b>	A	3
		B	4
<b>AI-written</b> article	correctly declared as <b>AI-written</b>	A	5
		B	6
	wrongly declared as <b>human-written</b>	A	7
		B	8

*Note.* Own illustration based on Graefe et al. (2018).

The experimental design was based on the procedure developed by Graefe et al. (2018), who previously investigated the perception of computer-written news pieces. Accordingly, the experiment was set up in a within-subjects design, with each test person passing through two of

the eight possible treatments shown in Table 3. Before entering the experimental setup, test persons were told that they are about to assess one human-written and one AI-generated article. All subjects were then presented with one article on each of the two topics in randomized order, whereby they were unaware about the real source of the articles. According to the alleged test setup provided to the test subjects, one article was declared as written by an AI and one as written by a human journalist (manipulation level 1), again in randomized order. The subjects thus always evaluated one article declared as AI and one declared as human. In addition, the real article source (AI or human) was randomly assigned (manipulation level 2). Thus, subjects could pass through the experimental setup reading two correctly declared, two wrongly declared or one correctly and one incorrectly declared article. This resulted in a total of 16 different possible experimental conditions, that subjects were randomly assigned to (see Table 4).

**Table 4**  
*Possible experimental conditions*

Article 1			Article 2			Condition	
Topic	Real source	Declared source	Topic	Real source	Declared source		
1	human	human	2	AI	AI	1	
		AI		human	human	2	
		human		AI	AI	3	
		AI		human	human	4	
	AI	human	2	AI	AI	AI	5
		AI		human	human	6	
		human		AI	human	7	
		AI		human	human	8	
2	human	human	1	AI	AI	9	
		AI		human	human	10	
		human		AI	human	11	
		AI		human	human	12	
	AI	human	1	AI	AI	AI	13
		AI		human	human	14	
		human		AI	human	15	
		AI		human	human	16	

*Note.* Participants were randomly assigned to one of the 16 experimental conditions.

Immediately after reading each of the two contributions, the text evaluation followed, whereby the test persons gave their rating of the respective text on a five-point Likert scale within the framework of predefined evaluation categories and corresponding items. The test persons evaluated the article

on the basis of the established categories ‘readability’, ‘journalistic expertise’, ‘credibility’ (Graefe et al., 2018) and ‘instrumental trust’ to rate the perceived text quality and the level of trust in the arguments and information presented. A detailed description of the operationalisation of the text evaluation categories can be found in Chapter 6.3. After going through the experimental procedure, the subjects were informed about the possibly misstated origin of the texts. Subjects were intentionally deceived by misinformation, about which they must be informed afterwards for ethical reasons (Dzeyk, 2001, pp. 15-16). In a short debriefing, the purpose of the research, their experimental condition and the actual source of the texts read were explained to the subjects (see Appendix 4). In this way, the participants also got a coherent impression of the experiment.

The data collected in this way resulted in 2x2x2 comparison groups that provide insights into the research questions and hypotheses (Bryman, 2012, p. 58). First, the average rating of texts declared as human and of those declared as AI-produced can be compared to each other, regardless of their actual source (regarding manipulation level 1). Second, the average rating of texts actually written by a human can be compared to the rating of actually AI-produced texts (manipulation level 2). Third, although it does not provide relevant insights regarding the research questions, it could be determined which topic performed better. Creating the experiment in the within-subjects design had the advantage of handling the available subjects more efficiently. Since all subjects pass through two conditions of the independent variables, the number of subjects required is reduced compared to the alternative between-subjects design. For each run of a block of the 16 conditions, ratings could thus be collected 16 times for actually AI or human-written texts as well as 16 times for texts with both source declarations (see Table 4). In the between-subjects design, by contrast, only one experimental condition would be assigned to each subject, meaning that barely enough data could have been collected per condition. This factor was especially important given the difficulty of recruiting subjects privately without using recruiting platforms or panels (Sedlmeier & Renkewitz, 2018, p. 157).

In order to be able to eliminate the influence of potential confounding factors (see Chapter 6.1) as much as possible, various measures were taken. As mentioned above, the texts were presented in a randomised order in all respects. Both the order of the actual source of the texts (human or AI) as well as of the topic (1 or 2) were randomly assigned (following Graefe et al., 2018). In both positions, each participant was assigned with equal probability (50%) to either an actual human text or an actual text written by an AI, as well as to either topic A or topic B. In addition, the declaration of the correct or incorrect text source was again randomised, which likewise resulted in a 50% probability of being assigned to each case. Thus, it can be assumed that all person-related

confounding factors are equally distributed across the eight treatments and consequently neutralised (Sedlmeier & Renkewitz, 2018, p. 141). The randomisation took place ex-ante, that is, before the start of the experimental procedure (Diekmann, 2005, p. 289) and was carried out via the applied survey software Unipark. The respondents were each assigned to one of the 16 possible article combinations (see Table 4) as part of a block randomisation. Blocks of 16 subjects were formed, to each of which one experimental run was randomly assigned one after the other (Sedlmeier & Renkewitz, 2018, p. 141). Technically, this was realised in Unipark by means of a coincidence trigger, which generates a random number between one and 16 for each test person. The number again represents the corresponding experimental run, which the participant is then automatically forwarded to. An equal distribution was aimed for in the drawing, which is equivalent to a so-called drawing without laying back (SoSciSurvey, 2019). Thus, each number can only be drawn once until all 16 possibilities have been exhausted. This procedure has the advantage that all experimental conditions are assigned the same number of times and a roughly equal and thus more comparable group size per condition is achieved.

Moreover, with regard to the variation of the independent variable declared source (human or AI), as well as topic (A or B), all person-related confounding variables are neutralised as effectively as possible in the applied within-subjects design anyway. Since each subject passes through all possible variations of these two variables, no confounding can occur (Sedlmeier & Renkewitz, 2018, p. 157). For example, a particularly critical person (personality-dependent confounding factor) would accordingly give a critical rating in both conditions of a variable and not just in one. Consequently, there can be no bias in the results, as possible effects occur in every condition.

Similarly, position effects that typically occur in within-subjects designs are ideally controlled. For instance, participants might get used to the task and thus change their behaviour when reading and evaluating the second text. Due to the repetition, the second text could either be evaluated more critically because of a practice effect and the possibility to compare to the first text, or less carefully because of fatigue or boredom. However, since the equal distribution of different treatments ensured that all options were passed through equally often at the same position (first or second), such positional effects were eliminated as confounding factors (Sedlmeier & Renkewitz, 2018, p. 162). As Table 4 shows, the different conditions were presented in a fully balanced manner (Sedlmeier & Renkewitz, 2018, p. 166), arranging all treatments systematically in every possible order (van der Lee et al., 2019, p. 8). Accordingly, any possible position effect should occur equally in all conditions again.

### 6.3 Operationalising of article perception

After the test subjects had read the respective article within the experimental setup, they were asked to rate its quality. The operationalizing of the article quality through different constructs and items as well as the design of the used rating scale will be explained in the following.

The measurement of the quality perception of the articles was based on a self-assessment of the subjects' explicit personal opinions towards the articles. Explicit attitudes here mean such attitudes that are directly expressed verbally (Garms-Homolová, 2020, p. 53). A self-report on the perception of the texts was used, which could be expressed on a rating scale (Döring & Bortz, 2016, p. 244). Specifically, the level of agreement (Döring & Bortz, 2016, p. 245) with individual statements about the text quality was queried. For this purpose, established and predefined question categories and items were used. The evaluation of the articles was based on the categories text quality and trust. The perceived *text quality* was assessed using the constructs *readability* (well-written, entertaining, interesting), *journalistic expertise* (coherent, comprehensive, descriptive) and *credibility* (fair, trustworthy, reliable) with three items each. These constructs and items are taken from the study by Graefe et al. (2018) and have proven to be useful for assessing journalistic text quality. They originally stem from the work of Sundar (1999), who investigated readers' criteria for news assessment. The evaluation categories were measured by the level of agreement with the corresponding statements (see Table 5).

**Table 5**  
*Operationalisation of article perception*

Construct	Item	Wording	Source
Readability	wellwritten	The article is formulated in a linguistically appealing way.	Graefe et al. (2018); Sundar (1999)
	entertaining	The article is written in an entertaining way.	
	interesting	The article is written in an interesting way.	
Journalistic expertise	coherent	The argumentation in the article is coherent.	
	comprehensive	The article presents the topic comprehensively.	
	descriptive	The article argues descriptively.	
Credibility (Intuitive trust)	fair	The article argues in a fair way.	
	trustworthy	The article generally seems trustworthy.	
	reliable	The article contains reliable information.	

*Operationalisation of article perception*

Instrumental trust	opinion	Based on this article, I can build an informed opinion on the subject.	Prochazka (2020)
	arguments	I would express the opinion presented in this article to friends and colleagues as well.	
	facts	I would use information from this article in conversations with friends and colleagues.	
	recommendation	I would recommend this article to a friend.	

*Note.* The wording of the items was presented to the subjects in German in the experiment.

In addition, the level of trust in the respective article was measured using the constructs intuitive trust (credibility) and instrumental trust (opinion\_building, arguments, facts, recommendation). This division was based on (Prochazka, 2020), who developed this operationalisation in the context of measuring trust in online journalism. The concrete formulation of the items was modified accordingly for the purpose of this work, so that the assessment of trust in the respective text was possible. Querying the intuitive trust aims at an initial, general and spontaneous assessment of the level of trust placed in the article (Prochazka, 2020, p. 168). The construct ‘credibility’ represents this exact same idea of intuitive trust. This is why they are both considered the same in this context and credibility is also part of the operationalisation of the level of trust in the text. The concept of the instrumental measurement of trust goes beyond the intuitive trust and additionally includes how recipients concretely deal with the information and arguments contained in the text. It is about whether readers in real life actually take the risk of trusting the content of the article. In concrete terms, this means the extent to which readers take the personal risk of deceiving themselves on the basis of the article, forming an inaccurate opinion or being considered uninformed in a social context (Prochazka, 2020, p. 168).

Accordingly, a total of 13 rating items were queried per article, which were divided into two item groups in the questionnaire (see Table 5). The first section contained the six items of the constructs readability and journalistic expertise and was titled ‘Evaluation of text quality’. The second section contained the seven items of the constructs ‘credibility’, ‘instrumental trust’ and the item ‘recommendation’ and was called ‘Evaluation of credibility’. The order of the items was randomised in each section to prevent bias due to order effects (Döring & Bortz, 2016, p. 254). The scale used was a fully verbalised, unipolar, five-point rating scale from ‘do not agree at all’ to ‘fully agree’, on which the subjects could express their agreement or disagreement with the respective statement. Five- to seven-item scales provide a solid degree of differentiation between response options and thus offer the best possible reliability and validity (Menold & Bogner, 2015, p. 2; Döring & Bortz,

2016, p. 249). Seven levels would make the question unnecessarily complex and possibly overwhelm some participants, as they would not be able to make such a precise judgement (Porst, 2014, p. 87). In line with the approach of Graefe et al. (2018), a five-point scale was used, as fewer points would give respondents too little flexibility to make a differentiated judgement (Döring & Bortz, 2016, p. 249; Porst, 2014, p. 94). Following the proposal of Döring & Bortz (2016), the verbal naming of the scale recommended by (Rohrmann, 2007) for querying agreement to statements in German translation was chosen. The options thus ranged from ‘fully disagree’ (‘stimme überhaupt nicht zu’), ‘mainly disagree’ (‘stimme eher nicht zu’), neutral, ‘mainly agree’ (‘stimme eher zu’) to ‘fully agree’ (‘stimme voll und ganz zu’) (Döring & Bortz, 2016, p. 245). The chosen terms suggest an equal distance between the answer options (Döring & Bortz, 2016, p. 245; Rohrmann, 2007). With regard to the statistical evaluation, this is necessary in order to fulfil the highest possible scale level and to achieve an interval scale level (Porst, 2014, p. 94). Interval scaling requires an equally perceived distances between the individual scale points (Döring & Bortz, 2016, p. 244; Porst, 2014, p. 75). This can be considered to be the case with the five-point scale chosen, as the options are named at equal distances, and the scale can therefore be considered to be interval-scaled (Döring & Bortz, 2016, p. 251). The scale was arranged horizontally from left (negative) to right (positive) in the questionnaire. Compared to a vertical presentation, respondents in this arrangement tend less to choose the extreme values at the edge due to possible primacy and recency effects (Tourangeau et al., 2000). However, in the mobile view on the smartphone, the responsive design of the questionnaire presented the response options one below the other, from negative (top) to positive (bottom).

The five-point scale thus also offers a middle option, which allows respondents to indicate a neutral attitude. In this way, it can be avoided that people with a neutral attitude towards the respective item select an inappropriate category and tend to agree due to social desirability, for instance. Agreement would therefore possibly occur more frequently on a scale without a centre (Menold & Bogner, 2015, p. 5). This could lead to biases in the data, while a middle category can increase the reliability of a survey (O’Muircheartaigh et al., 2000). This advantage is opposed by the fact that the middle category may at the same time intensify the trend toward choosing the neutral middle (Döring & Bortz, 2016, p. 249). Especially tired and unmotivated people can be encouraged to choose the middle more often in order to reduce their own effort. Nevertheless, it is commonly recommended to offer a medium category in order to prevent evasion to wrong answers (Menold & Bogner, 2015, p. 5). Since especially in the case of the text evaluation at hand it can be assumed that a neutral attitude towards a quality feature of a contribution occurs quite often, a middle category was offered. The design of the scales was kept the same throughout the experimental

procedure in order to remain consistent and make the questionnaire easier to understand (Porst, 2014, p. 95).

## 6.4 Operationalising of personal attitudes towards AI

The first part of the online survey focused on personal attitudes towards the topic of artificial intelligence, prior to the experimental setting. After a short introduction explaining the aim of the study, the first part of the survey asked about the subjects' personal attitudes towards the topic of artificial intelligence. On the one hand, this is necessary against the background of RQ 3, which examines the influence of personal attitudes on the perception of AI-generated texts. On the other hand, the questions in combination with the opening text served as an introduction to the topic. The range of personal attitudes towards AI was assessed using the constructs 'previous experience with AI', 'general trust' and 'benevolence', as well as 'trust in the capacities of AI' (see Table 6). These personal opinions towards the topic of artificial intelligence were again queried by means of a self-assessment of the explicit attitudes of the test persons (Döring & Bortz, 2016, p. 244), as with the evaluation of the texts.

**Table 6**

*Operationalisation of personal attitudes towards artificial intelligence (AI)*

Construct	Item	Wording	Source
Previous experience	knowledge	Have you ever read or heard of the term 'artificial intelligence'?	Bitkom Research (2017)
Trust	trust_personal	I would have no difficulty trusting an artificial intelligence with personal matters.	Wullenkord (2017)
	trust_ethics	I trust artificial intelligence to always act ethically.	-
Benevolence	intentions	Artificial intelligence is being developed with good intentions.	Chien et al. (2014)
	chances	Artificial intelligence generally offers more opportunities than threats.	Chien et al. (2014)
Trust in capacities	tasks	Artificial intelligence can solve tasks just as well as a highly qualified human being.	Chien et al. (2014)
	reasoning	Artificial intelligence is able to argue just as logically as a human being.	Wullenkord (2017)

*Note.* The wording of the items was presented to the subjects in German in the experiment.

Since the first questions are crucial for the successful course and completion of the survey, they were asked in a personal, current and interesting manner (Porst, 2014, p. 142). The personal



background, previous experience and attitude towards the topic of artificial intelligence were explicitly targeted in order to arouse the respondents' interest, increase their personal involvement and thus prevent immediate dropouts. At the same time, care was taken not to make the initial questions too complex, but as easy as possible for everyone to answer. A first question on previous knowledge, in which respondents could indicate their personal level of knowledge about the topic of AI, served as an introduction to the questionnaire. The question 'Have you ever read or heard of the term artificial intelligence?' was presented separately from the subsequent items. There was a choice between the answer options 'No, I have never heard of it', 'Yes, but I don't know exactly what it is', 'Yes, and I know roughly what it means', 'Yes, and I can explain well what it means' and 'Yes, and I have expert knowledge'.

In a set of items, the dimensions of general trust, benevolence and trust in capacities were then asked, which were focused on personal attitudes towards AI. The selection of the items was based on existing research on comparable topics and appropriate items were adapted. All items were formulated positively in order to avoid confusion through complicated negations (Porst, 2014, p. 107). In addition, care was taken to use wording that was as direct, simple and universally understandable as possible. The items were presented in random order. Similar to the text perception questionnaire, a fully verbalised, unipolar, five-point rating scale from 'strongly disagree' to 'strongly agree' was used in order to keep the questionnaire as consistent as possible (see Chapter 6.3).

## 6.5 Generation of articles

In the following, the selection of the articles as well as the procedure for generating the AI-written versions will be discussed. The settings used for the available parameters of the GPT-3 language model will also be discussed.

One human and one AI-written journalistic opinion piece on two different topics were required as treatments. In other words, a total of four different texts were used, which presented an opinion on a well-known and generally understandable topic from recent times in German. Both the human and AI-written contribution per topic argued in the same direction in order to keep them comparable. In addition, a comparable text length was ensured, roughly following Tewari et al. (2021), who suggested a text length of about 400 words in their experiment on AI-written news. This ensured that the texts can be read in a reasonable amount of time (Tewari et al., 2021, p. 4). To avoid personal involvement by the test persons, very controversial and sensitive topics as well as AI outputs with extreme opinions were excluded. In addition, it was made sure that the articles were not directly related to current issues, but that the corresponding matter was already covered

a few weeks ago. To generate the AI texts, various suitable human opinion articles on current topics were first retrieved. The beginnings of the articles were used as input and given to GPT-3 as a prompt to complete the text. Thus, some inputs were experimented with and several AI-generated results were compared to each other to select two articles (procedure following Tewari et al., p. 4). Ultimately, it was decided to use the two human articles (Text 1 & Text 2) which allowed the AI to generate the best linguistic output. Text 1 appeared on *tagesschau.de* and dealt with the commemoration of the terrorist attack on the 1972 Olympic Games in Munich and the resulting discussion about compensation payments for relatives (see Appendix 3.1). Text 2 is a commentary on *deutschlandfunk.de* which discusses the death of the Queen of Great Britain and the impact on the country (see Appendix 3.3). In both texts, the names of the authors were changed and the dates removed in order to produce the human text version for the experiment. Text 1 contained two subheadings, which were removed to make both texts more comparable. In text 2, the third paragraph of the human original mentioned the former British Prime Minister Liz Truss, a person who was no longer in office at the time the experiment was published. This was not predictable when the texts were selected. Because the article appeared a few weeks before the study was conducted, an inconsistency arose here that would certainly have been noticed by some readers. This could have led to an unjustified, worse evaluation of the human text in this case, which is why the corresponding passage was slightly modified. Apart from that, nothing had been adjusted. The two human contributions including the minor adjustments are also provided in Appendix 3. They served as comparison and control variables in the experimental set-up and at the same time as an input for generating the AI-written texts.

To generate the texts, GPT-3 was first given a concise instruction (prompt) to complete the following opinion piece. Furthermore, it was instructed to ensure a good linguistic style and logical argumentation. This was followed by the actual content input, which was taken from the real journalistic opinion pieces and consists of the headline, the subtitle, the lead and the first paragraph of the original article. It was also made sure that the input ended with the first three to four words of the sentence that followed. Having an incomplete sentence completed in this way and defining it as the starting point for the AI is a common practice (Brown et al., 2020; Gunser et al., 2022; Tewari et al. 2021) and gave the best results during the tests of GPT-3. The exact part of the text that served as an initial input to GPT-3 is indicated in the AI-generated texts in Appendix 3.2 and 3.4. The text input was made available to the AI via the Playground interface (OpenAI, n.d.-e). The AI can be launched with the given input and generates a suitable output. In further attempts, output can then be generated again from any point in the text. For instance, further output can be generated from the end of the previous output or from possibly incorrect text parts. Since GPT-3

does not always generate the same output for one and the same input, different results with different quality can be obtained when repeating the procedure several times. Therefore, it had to be ensured that not arbitrarily many trials were carried out until the quality met eventual expectations of the user. Based on experience from previous tests with GPT-3, guidelines for the production of the text were derived to ensure that the generation process was always the same and the output produced under the same conditions.

Thus, GPT-3 received the same number of maximum attempts to correct itself in each text in the case of errors to make for a comparable amount of human assistance. Once an obvious error of a linguistic or factual nature was generated by the AI in a section of text, that section was deleted. The AI was then given a maximum of five new chances to continue the text in a meaningful way. If incorrect passages occurred again in the further course of the text, the AI was again given a maximum of five opportunities to improve the text until it was brought to a coherent conclusion. This procedure was carried out for a total of 10 different topics respectively inputs and the two most coherent completions were selected, similar to the approach of Tewari et al. (2021, p. 4). Editing was kept to an absolute minimum. Minor inconsistencies or linguistically suboptimal formulations were kept in order to ensure the authenticity of the text. All articles are provided in Appendix 3. In order to vary the treatments, each of the two AI-written as well as human-written articles were created with both possible source declaration, making for one correctly declared and one incorrectly declared source (AI or human-written). For this purpose, the recommendation for attributions to computer-written texts by Montal & Reich (2017) was roughly followed. For the AI source attribution, the name of the AI and the developer were mentioned in a typical byline before the actual text. In an extended disclosure at the end of the text, GPT-3 was again mentioned as a product of OpenAI and the machine origin of the text was pointed out (Montal & Reich, 2017, p. 15).

As suggested by OpenAI themselves, all text examples were generated using the davinci (*text-davinci-003*) model as it is the most advanced language model of GPT-3. Accordingly, it promises the best outcomes compared to the other models. In addition, OpenAI explicitly states complex issues, logic and an understanding of cause and effect as davinci's strengths (OpenAI, n.d.-d). Since these are relevant skills when writing an opinion piece, davinci was suited best for the application at hand. Besides, GPT-3 offers various adjustment parameters that have an effect on the quality of the results. A detailed explanation of the parameters and their effects on the output was given in Chapter 3.2. In order to determine the best performing values for the present use case, several test runs were carried out. The temperature option was used as the sampling method to select a word

(token) from the predicted amount of possible next words. As recommended, the alternative top\_p sampling value was set to the default value of 1, as only one of the two sampling methods should be used. The temperature value was set to 0.7, which ensured a sufficient degree of creativity and variation. In addition, the presence penalty was set to 0.4 and the frequency penalty to 0.6 to encourage linguistic and thematic twists and reduce repetitive results, such as the repetition of words and lines. The maximum response length was set to 2800 tokens, which corresponds to about 700 characters.

## 6.6 Pretest of the experimental setup

A pre-test was carried out to check the comprehensibility and the correct technical operation and procedure of the experimental set-up. The minor adjustments that were made on the basis of the findings from the test are presented below.

The pre-test first included a mere technical check of the questionnaire. Test data was automatically generated using Unipark in order to test the general functionality of the survey and the routing of the individual test persons through the different experimental conditions. For this purpose, the filter function created for randomisation and the equal distribution of the test persons to the 16 different experimental conditions were tested. Apart from that, three test persons completed a test run. They were asked to give their thoughts on inconsistencies and other considerations via the pretest comment function in Unipark. The test thus took place under real conditions, again paying attention to the correct filtering. Based on the feedback received in this way, a few wordings were revised in the welcoming, in the instructions and in the debriefing at the end. In the pretest, four items were asked per rating dimension to evaluate the text. The item that turned out to be the most unclear and most difficult for the test persons to answer was removed in each case. This resulted in the three items actually used per rating dimension (see Chapter 6.3). It was also found that adjustments were necessary for the mobile use of the survey on smartphones. Originally, the Likert scale for rating the items was designed as a scale where only the endpoints were named. Since the arrangement of the answer options is adapted to the screen sizes of the respective end devices by the responsive design of the questionnaire, the options were arranged one below the other in the mobile view. This meant for one that the character of the Likert scale was lost. For another, the scale was no longer clearly understandable, as the lack of labelling resulted in empty lines below each other. Accordingly, the only end-point named scale had to be converted into a fully verbalised scale for mobile use, so that the answer options would be clear. In addition, the articles were not inserted as an image file as originally planned, but in text form so that they remain readable on the

smartphone. The time required by the respondents in the pretest, including comments, was between 11 and 12 minutes.

## 6.7 Data collection and sample

The survey period began on the 24th of October 2022 and ended on the 20th of November 2022. After data cleaning and the exclusion of incomplete responses, 192 participants remained. The participants for the study were recruited among one's own circle of friends and colleagues in the professional and academic environment. The experiment was shared online via platforms such as WhatsApp and Instagram. In addition, the study was distributed in mailing lists of the University of Applied Sciences Darmstadt and the University of Mainz. Furthermore, the experiment was shared via listings on the survey-sharing services SurveySwap and SurveyCircle and on the site [psychologie-heute.de](https://psychologie-heute.de). In this way, people with different backgrounds could be recruited, both from the extended personal environment and strangers.

The sample is a non-probabilistic convenience sample (Döring & Bortz, 2016, p. 196). The experiment was shared via the internet and was in principle available to all internet users. All subjects who decided to participate were accepted. The test persons were therefore selected at will, which is why it is unknown whether and to what extent the composition of the sample differs from the population (Döring & Bortz, 2016, p. 305). This arbitrarily drawn sample among all internet users is suitable for testing the presumed effects of the independent on the dependent variables within the framework of the experimental design, but does not allow any conclusions to be drawn about the German-speaking population. As a study conducted online, the sample scope is limited to German-speaking internet users (Hauptmanns, 1999). People who do not use the internet are not accessible. Furthermore, not all classes and groups of the population can be found online (Gnambs & Strassnig, 2007, p. 237). This means that an equal distribution of social and demographic backgrounds of the test persons is not guaranteed and the sample is not representative of the of the German-speaking population. This is also reflected in the demographic data of the sample. Due to the environment in which the study was conducted, the sample is young with a high level of education. Of the 192 participants, 42% identified themselves as male, 56% as female and 2% as diverse. While 79% are under 34 years old, only 21% are 35 or older. 32% of the respondents have a higher education entrance qualification and 54% have completed a higher education degree (34% bachelor's degree, 17% master's degree, 3% PhD). Only 14%, on the other hand, have not completed a degree. Regarding the optimal sample size for experiments that test two variables in two levels within-subjects, Brysbaert (2019, p. 27) suggests about 145 subjects as a minimum. Considering the clearly higher number of 192 participants and the fact the research

interest is concentrated on the variation of two variables (actual and declared source), the sample size can be described as sufficient.

## 6.8 Procedure for data evaluation

Before conducting the actual data analysis, the raw data had to be correctly prepared and cleaned. Furthermore, the reliability of the used scales was calculated and the correct randomization as well as certain requirements for the statistical analysis were checked. The process is described in this chapter.

Out of a total of 262 responses, 65 incomplete questionnaires were excluded. This resulted in 197 subjects who completed the entire experiment. In addition, five subjects were excluded who completed the experiment significantly too fast at under 170 seconds. This left a total of 192 data sets for evaluation. The average completion time was 12 minutes and 55 seconds. The raw data were downloaded from the used survey platform Unipark as an Excel file. In order to verify the quality of the data before the evaluation, a successful randomisation was checked first. Of a total of 262 participants, 27 dropped out even before they entered the experimental setting and were assigned to a condition. This left 235 participants who were assigned to an experimental condition. Overall, each condition was assigned with the same probability of about 6%, which corresponds to an optimally randomised assignment (for 16 conditions). In addition, possible systematic dropout was checked for. As already mentioned in Chapter 6.2.1, systematically accumulated dropouts of participants in certain experimental conditions may threaten the correct randomisation and neutralisation of confounding variables. In total, 38 dropouts were distributed across all conditions, with all conditions being discontinued between one time to four times, and conditions no. 4 and no. 13 recording no dropouts. Thus, all 16 experimental runs were completed between 10 and 14 times. Accordingly, no highly concentrated dropout can be seen in any of the conditions. Accordingly, each of the 16 different conditions was still accomplished between 6% and 7% even after removing the dropped-out questionnaires. This resulted in approximately equal group sizes per experimental condition, as Table 7 shows.

**Table 7***Group sizes for individual experimental conditions*

		Real source		
		Human	AI	Total
Declared source	Human	98	94	192
	AI	100	92	192
	Total	198	186	

*Note.* The group sizes result from random group assignment.

Afterwards, the data was prepared in Excel for evaluation. The varied independent variables ‘real source’ and ‘declared source’ were each coded with ‘1’ for ‘human’ and ‘2’ for ‘AI’. The independent variable ‘topic’ was also coded with ‘1’ (Topic 1: Olympics) and ‘2’ (Topic 2: Queen). The independent variables coded in this way could then be used to create the descriptive charts for the results in the different experimental conditions and for the analysis of variance.

Subsequently, the reliability of the applied scales was tested. For the scale of ‘attitudes towards AI’, a relatively low Cronbach’s alpha value of  $\alpha = .61$  was found, which represents a weak but still acceptable reliability (Bryman, 2012, p. 170). All items were combined into one scale for the attitudes. The item ‘previous knowledge’ was evaluated individually, as it was an introductory question with concrete answer options that did not ask for agreement with statements and is therefore not comparable with the other items. The scales for the text ratings ‘credibility’ ( $\alpha = .77$ ), ‘readability’ ( $\alpha = .79$ ), ‘expertise’ ( $\alpha = .78$ ) and ‘instrumental trust’ ( $\alpha = .84$ ) showed high reliability. The mean values were calculated from all underlying items for each scale or rating dimension (Huber, et al., 2014, p. 49) prior to evaluating the descriptive results.

In order to answer the research questions and test the hypotheses, a multi-factorial, multivariate analysis of variance (multi-factorial MANOVA) was carried out using SPSS. This was used to determine whether possible mean differences could be attributed to the independent variables. The factors actual source, declared source and topic, which varied in the experimental design, were defined as independent variables (multi-factorial) and the dependent variables readability, expertise, credibility and instrumental trust were defined as dependent variables (multivariate). Since applying a MANOVA requires certain preconditions, these were checked before the actual evaluation, following Huber et al. (2014). There should be no strong outliers in the data and the dependent variables should be linearly correlated (Huber et al., 2014, pp. 63-69). An exploratory data analysis generally did not reveal any significant outliers in the data. Due to the closed design of the applied

5-point scale, however, outliers are not likely anyway. The test for linear correlation of the dependent variables showed that they are sufficiently correlated with each other. With values between  $r = .55$  and  $r = .72$ , they lie around the guideline value of  $r = .6$ . A graphical examination also revealed a correlation between all variables. At the same time, no excessively high correlation (multicollinearity) was found between the dependent variables. All correlations remained below the critical value of  $r = .9$ . (Huber et al., 2014, pp. 68-69). Normal distribution (univariate and multivariate) of the data is not given, yet this does not matter due to the approximately equal group sizes in all experimental conditions (see Table 7). The largest and smallest group size differ by considerably less than the critical factor 1.5, which is why equal group sizes can be assumed (Huber et al., 2014, pp. 64-65).

A MANOVA could therefore be used to identify possible significant effects and was conducted using SPSS. The results of the MANOVA (see Appendix 2) were then examined with regard to significant effects of the independent variables in order to test the established hypotheses. With regard to RQ3, the correlation between the attitudes towards AI and the ratings in the relevant conditions was also to be examined. For this purpose, the Pearson correlation coefficient was calculated between the variable 'Attitudes towards AI' and the rating values of the constructs 'readability', 'expertise', 'credibility' and 'instrumental trust', in each case for both conditions of the declared source (AI & human). For this purpose, the data set was filtered in SPSS for declared source 'AI' (2) and declared source 'human' (1) before calculating the correlations for each. Afterwards, the correlation values calculated were compared to determine whether there were differences in the relationship between personal attitudes and the ratings of AI or human-written articles. Since correlation coefficients ( $r$ ) are not metrically scaled, meaning that they are not equidistant, different correlations cannot be directly compared to each other. For example, double values do not imply a correlation that is twice as strong (Bortz, 2005, p. 219). Therefore, a transformation of the  $r$ -values into the metric Fisher Z-values was necessary in order to obtain standardised and thus comparable values. The values for this were calculated according to Bortz (2005, p. 219). The Fisher Z-values listed in Table 8 in the results chapter make a direct comparison possible between the individual correlations in the two conditions. Using the mean value of the Z-values per condition, the mean  $r$  of the overall ratings was then determined in each case, which is thus also in a comparable relationship to each other.



## 7 Results

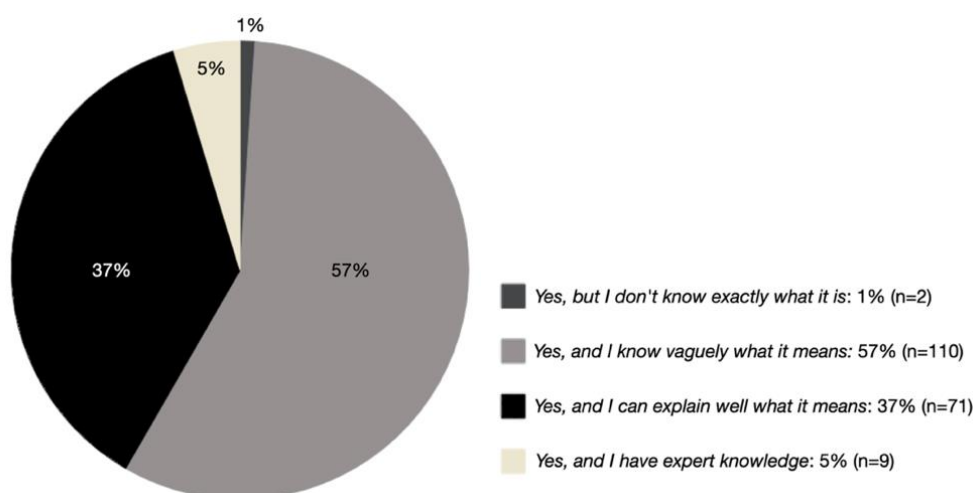
In the following, the results of the conducted experiment will be depicted. Descriptive findings of the assessment of articles in the different experimental conditions will be presented first. The results are then analysed with regard to the research questions. Using the results of a multivariate analysis of variance (MANOVA), significant effects of the independent variables (actual and declared source) are identified to test the hypotheses. In the concluding section, all results are then discussed in terms of their implications and in the context of the current state of research, as well as evaluated with regard to their limitations and quality criteria.

### 7.1 Descriptive findings

As a first overview, the results on the participants' attitudes towards AI and their previous experience with the topic are presented below. Subsequently, the descriptive results of the actual experiment are presented by showing the article ratings for the individual test conditions. The mean of the general attitudes towards AI (composed of 'trust in AI', 'benevolence' and 'trust in capacities' of AI; see Chapter 6.4), comes to 2.95 ( $SD = .6$ ). This value represents an almost exactly neutral or mid-point attitude of the entire sample on the five-point Likert scale.

**Figure 2**

*Previous knowledge regarding the term 'artificial intelligence'*



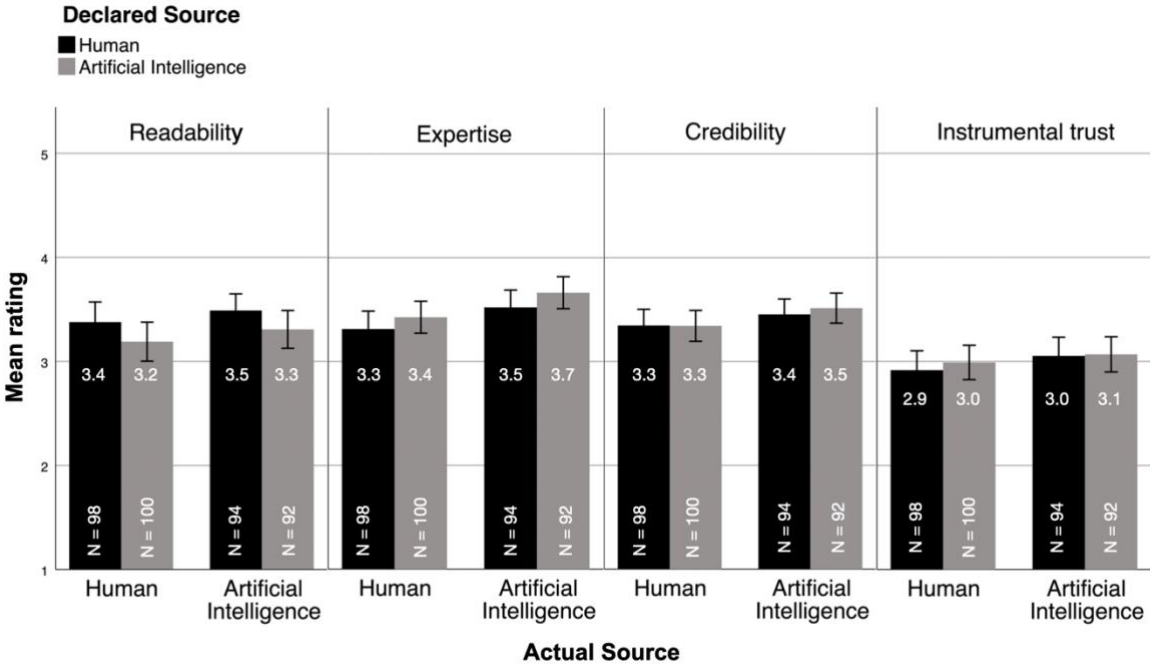
*Note.* Answers to the question: „Have you ever read or heard of the term 'artificial intelligence'?"

In addition, a single item was used to measure previous experience or knowledge on the topic, which is shown in Figure 2. For this purpose, the following question was asked: 'Have you ever

read or heard of the term artificial intelligence?’ A majority of 57% stated that they knew roughly what the term artificial intelligence meant. 37% were confident that they could explain well what it means and 5% said they had expert knowledge. Only 1% of all participants said they did not know exactly what it was. No one stated they had never heard of the term ‘artificial intelligence’.

Regarding the results of the actual experiment, Figure 3 shows the overall mean ratings on the five-point scale of all articles including both topics, divided by their actual source and the declared source. It becomes evident that actually AI-written texts tend to be rated slightly better than human versions for all the variables readability, expertise, credibility and instrumental trust. No matter which source was declared, the actually AI-written articles were rated slightly better in any case.

**Figure 3**  
*Mean article ratings per actual and declared source combined*



*Note.* Error bars indicate 95% confidence intervals.

Furthermore, correctly declaring AI-written texts as such had a positive effect on ratings in almost all criteria. Even declaring the actually human-written articles as AI-written resulted in slightly higher ratings for most of the quality criteria. Only for readability, the declaration as AI-written resulted in slightly worse values than for stating a human journalist as text source, for both actual sources. The confidence intervals show the range in which the mean values would lie with a probability of 95% for repeated measurements. The intervals all overlap for the different experimental conditions. Thus, there are no clear differences with statistical significance in the

perception of the article versions. Overall, the variable ‘instrumental trust’ of all texts was rated noticeably worse than the three remaining variables ‘readability’, ‘expertise’ and ‘credibility’, which had roughly equal scores.

With regard to the two topics of the articles (Topic 1: Compensation for the 1972 Munich Olympics attack and Topic 2: Death of the Queen), the ratings did not differ on average. The mean values of the ratings in total per topic differ only for the variable readability. For readability, the articles on the Olympic Games attack (Topic 1) were thus rated worse ( $M = 3.2, SD = .87$ ) than the articles on the death of the Queen (Topic 2) ( $M = 3.5, SD = .92$ ). The remaining evaluation criteria showed the same mean ratings for both topics regarding expertise, credibility and instrumental trust. All mean article ratings across all conditions are listed in Appendix 1. Varying the variable ‘topic’ had a significant main effect on the readability,  $F(1, 376) = 5.855, p = .016, \eta^2_{\text{part}} = .015$ . The differences in perceived readability are thus attributable to the variation of the topic. However, the effects of the topic variable only play a secondary role here, as they do not provide any relevant insight regarding the research objective and hypotheses. Furthermore, there were no significant interaction effects of the independent variables (see Appendix 2).

## 7.2 Answering research questions

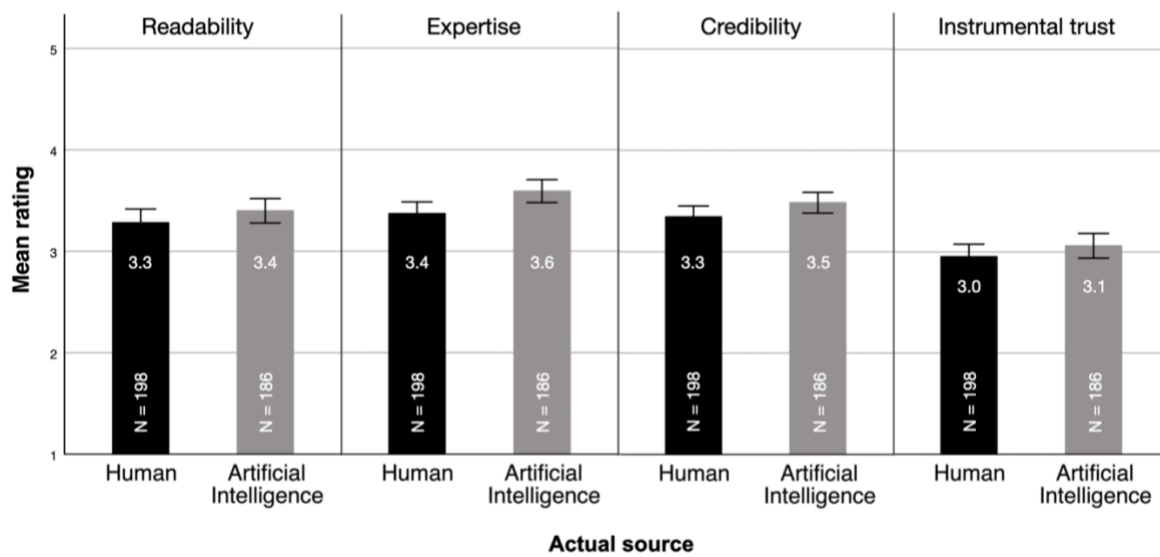
After an overview of the ratings of the articles for the individual experimental conditions was given in the previous chapter, statistically significant effects of the independent variables (real and declared source) will now be tested in order to answer the research questions.

RQ1 considered the effect of the actual source, thus asking ‘What influence does the actual source of the articles (AI or human) have on the perception of journalistic text quality and trust in the text?’ The question is answered and the corresponding hypotheses are tested on the basis of the mean ratings per actual source (see Figure 4) and the results of the MANOVA (see Appendix 2).

It was hypothesized, for one, that actual AI-written articles are evaluated more negatively than the human-declared text in terms of overall journalistic text quality (*H1.1*). For another, that actual AI-written articles are perceived more negatively than the actual human versions in terms of trust in the content (*H1.2*).

**Figure 4**

*Mean article ratings per actual article source*

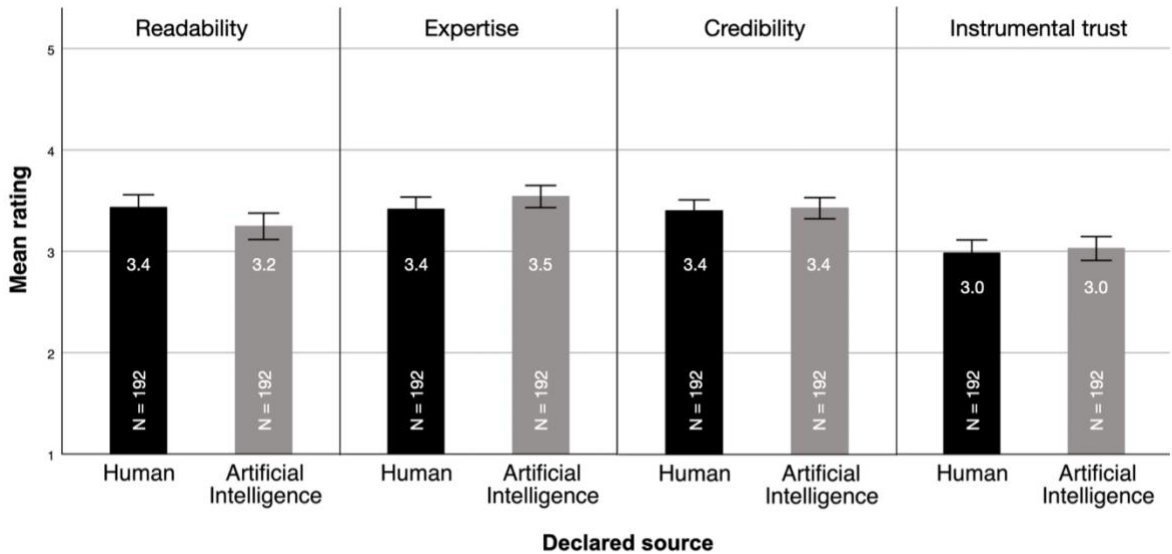


*Note.* Error bars indicate 95% confidence intervals.

Journalistic text quality (*H1.1*) was operationalised via the variables ‘readability’, ‘journalistic expertise’ and ‘credibility’. Trust in the articles is represented by ‘credibility’ (‘intuitive trust’) and ‘instrumental trust’. In terms of text quality, it is evident that articles written by AI were rated slightly better overall than articles written by humans. Regarding readability, AI-written articles ( $M = 3.4$ ,  $SD = .83$ ) were rated better on average than the human-written versions ( $M = 3.3$ ,  $SD = .96$ ). In terms of expertise, AI texts ( $M = 3.6$ ,  $SD = .78$ ) were rated better than human texts ( $M = 3.4$ ,  $SD = .86$ ). The credibility of AI-written texts ( $M = 3.5$ ,  $SD = .71$ ) was higher than for human-written articles ( $M = 3.3$ ,  $SD = .76$ ). In terms of trust placed in the content (*H1.2*), AI-written articles performed slightly better as well. As described, the credibility (‘intuitive trust’) of AI-written articles was rated higher than for human written texts. Regarding ‘instrumental trust’ AI-written texts ( $M = 3.1$ ,  $SD = .84$ ) were rated better than human-written ones ( $M = 3.0$ ,  $SD = .88$ ). AI-written articles from GPT-3 were therefore rated slightly better than the human versions in terms of journalistic text quality as well as trust. However, the differences are not significant as the confidence intervals overlap (see Figure 4). **Both H1.1 and H2.1 are therefore rejected.** Furthermore, the conducted MANOVA revealed no significant main effect of the independent variable ‘actual source’ on the overall rating of the articles,  $F(4, 373) = 2.219$ ,  $p = .066$ ,  $\eta^2_{\text{part}} = .023$ . Accordingly, varying the actual source (human or AI) has no significant influence on the rating of the four dependent variables and thus neither on the overall journalistic text quality, nor on the trust in the content.

RQ2 focused on the effect of source declaration, thus asking ‘What influence does the declared source of the articles (AI or human) have on the perception of journalistic text quality and trust in the text?’ The corresponding hypotheses are tested on the basis of the mean ratings per declared source (see Figure 5) and the results of the MANOVA (see Appendix 2).

**Figure 5**  
*Mean article ratings per declared article source*



*Note.* Error bars indicate 95% confidence intervals.

On one hand, it was assumed that ‘texts declared as AI-written are evaluated more negatively than the human-declared text in terms of overall journalistic text quality’ (H2.1). Text quality was operationalised via the variables readability, journalistic expertise and credibility. The descriptive results show that the readability of texts declared as written by an AI ( $M = 3.2, SD = .91$ ) was rated worse than texts declared as human-written ( $M = 3.4, SD = .88$ ). By contrast, the expertise rating shows a value in favour of the articles declared as AI-written ( $M = 3.5, SD = .77$ ), compared to those declared as written by humans ( $M = 3.4, SD = .84$ ). As for credibility, both source declarations score the same ( $M = 3.4, SD_{human} = .75; SD_{AI} = .73$ ). Accordingly, only the readability of allegedly AI-written articles turns out to be perceived slightly worse than allegedly human-written texts. The conducted MANOVA revealed a significant main effect of the independent variable ‘declared source’ on the overall rating of the articles,  $F(4, 373) = 3.769, p = .005, \eta^2_{part} = .039$ . The variation of the declared source (human or AI) thus generally influences the perception of the articles. Looking at the individual dependent variables (univariate), the declared source only shows a significant effect on the readability,  $F(1, 376) = 3.88, p = .05, \eta^2_{part} = .01$ . According to Cohen (1988, pp. 285-286), this main effect of the declared source is to be considered small ( $\eta^2_{part}$

= .01). This means that the slightly better rated readability in supposedly human-written articles is actually due to the variation in the declared source. However, the actual relevance of this effect outside of the given sample is questionable. As Figure 4 shows, the confidence intervals of both ratings of readability overlap, indicating no significant difference between both conditions (declared source human or AI). In terms of readability, there is thus only a very small negative effect at the expense of articles declared as AI-written. No significant effect on the other variables expertise and credibility were found (see Appendix 2). **Hypothesis H2.1 is therefore rejected.**

On the other hand, the hypothesis was formulated, that texts declared as AI-written are evaluated more negatively than texts declared as human-written in terms of general trust in the content (H2.2). Trust in the texts was operationalized by the constructs ‘instrumental trust’ and credibility (‘intuitive trust’). Both variables were rated exactly the same on average for each of the possible sources (AI or human). ‘Instrumental trust’ was rated in both cases with a mean of 3.0 ( $SD_{\text{human}} = .9$ ;  $SD_{\text{AI}} = .82$ ) and credibility (‘intuitive trust’) with a mean of 3.4 ( $SD_{\text{human}} = .75$ ;  $SD_{\text{AI}} = .73$ ). Accordingly, no significant effect of the declared source on the dependent variables instrumental trust or credibility were found (see Appendix 2). Texts declared as AI-written were therefore not rated worse in terms of trust placed in them than texts declared as human-written. **H2.2 is therefore also rejected.** With respect to RQ2, it can be concluded that there is no significant influence of the declared source (AI or human). The declared source had no relevant influence on the overall quality perception of the articles, apart from the small effect on readability.

RQ3 was concerned with the attitudes towards AI of the subjects and addressed the question of ‘what role the personal attitude towards AI plays when perceiving and evaluating AI-generated texts?’ It was assumed that ‘the attitudes towards AI are correlated with the perception of articles declared as AI-written.’ (H3.1), while at the same time, they are not correlated with the ratings of articles declared as human-written (H3.2). The correlations between the variable ‘attitudes towards AI’ and the rating criteria, divided by the declared source (AI or human), are shown in Table 8. For the case of articles declared as AI-written, there are slight to moderate significant positive correlations (Cohen, 1988, pp. 79-80; Ellis, 2010, p. 41) with the surveyed attitudes towards the topic of AI for each evaluation criterion. Personal attitudes correlate slightly with the rating of perceived readability,  $r(190) = .21, p < .01$  as well as perceived journalistic expertise,  $r(190) = .25, p < .01$ . The strongest correlation is found between attitudes and the credibility (‘intuitive trust’) with  $r(190) = .31, p < .01$ , as well as attitudes and instrumental trust  $r(190) = .32, p < .01$ . **H3.1 is therefore supported.** Meanwhile, the texts declared as human do not show a significant

correlation with the variable attitudes towards AI for any of the evaluation criteria. **H3.2 is thus also confirmed.**

**Table 8**

*Correlations between personal attitudes towards AI and article perception*

	<i>Readability</i>		<i>Expertise</i>		<i>Credibility</i>		<i>Instrumental trust</i>		<i>Mean</i>		
	AI	H	AI	H	AI	H	AI	H	AI	H	
<i>Declared source</i>											
<i>Attitudes</i>	<i>r</i>	.21**	.11	.25**	.05	.31**	.11	.32**	.11	.27	.01
	<i>Fisher-Z</i>	.21	.11	.26	.05	.32	.11	.33	.11	.28	.01

*Note.* Correlations between personal attitudes towards the topic of AI and the ratings for all quality criteria, for both possible declared sources ('AI' means artificial intelligence, 'H' means human). Fisher-Z values are directly comparable to each other.  $N = 192$ . \*\* indicates  $p < .01$ .

As different correlation coefficients of  $r$  are not directly comparable to each other, the values were additionally transformed into metrically scaled Fisher Z-values (Bortz, 2005, p. 219). The Z-values listed in Table 8 make a direct comparison between the individual correlations of the two conditions possible. It becomes evident again that all four evaluation criteria for allegedly AI-written texts show a stronger correlation with the attitudes towards AI.

Using the mean of the Z-values of all rating criteria, the mean  $r$  of the overall ratings was then determined for each declared source, which thus also stand in a comparable relationship to each other (Bortz, 2005, p. 219). Thus, the personal attitudes and the overall ratings of the allegedly AI-written texts correlate with  $r(190) = .27, p < .01$ , which according to Cohen (1988) nearly represents a medium correlation. The correlation of the attitudes with the mean ratings of the allegedly human-written articles is  $r(190) = .01$ , which can barely be interpreted as a small correlation. The ratings of allegedly AI-written articles thus show a stronger correlation with the attitudes towards AI than the ratings of allegedly human texts.

## *IV Concluding remarks*

### 8 Discussion

After having presented the results with respect to the research questions, the implications of the findings are discussed and placed in relation to previous findings. Subsequently, the methodological implementation of the experiment is discussed with regard to the quality criteria and general limitations.

#### 8.1 Implications of the results

The research has shown that journalistic opinion pieces written by the AI GPT-3 could keep up with texts written by a human journalist in terms of reader perception. Not only the journalistic text quality as such, but also explicitly the trust in the content and the argumentation were perceived as qualitatively equal (RQ1). All four evaluation criteria (readability, expertise, credibility/intuitive trust, instrumental trust) were rated even slightly better for the AI-written article versions. However, differences were not significant (see Figure 4).

The frequently postulated skepticism of readers towards allegedly AI-written texts could not be verified either. The survey of attitudes towards AI shows that the sample overall shows a rather neutral attitude towards AI, its possibilities and the trust placed in it ( $M = 2.9$ ,  $SD = .6$ ). Accordingly, no negative effect on the perception of articles could be found if the readers assumed that they were reading a text produced by an AI. Articles declared as AI-written were always rated equally well or minimally better than texts declared as human (RQ2). Only for the readability factor was there a very small, negative effect if the AI was declared as a source of the articles. However, this did not have a negative effect on the remaining evaluation criteria, such as trust in the content. These differences in the declared source were also minimal and not significant (see Figure 5).

For the topics at hand, the AI has produced an article quality perceived equal to human writing qualities in terms of all evaluation criteria, which underlines the enormous potential of AI-powered NLG solutions like GPT-3. These findings are surprising in so far as the texts written by the AI GPT-3 were perceived more positively than had been assumed on the basis of previous research results on AI texts. It turns out that AI-written articles containing own opinions and arguments can not only keep up with human versions in all respects, but mostly even slightly surpass them



(see Figure 3). These results imply that readers would hardly be able to distinguish between AI-written and human articles.

The results of this work are in line with previous findings regarding the perception of AI-written journalistic articles in particular. Brown et al. (2020) and Clark et al. (2021) already showed that readers could not reliably distinguish news written by GPT-3 in English from human versions. Even though an exact quality rating was not queried in these studies, these results indicate a human-like text quality of the GPT-3 texts, which is consistent with the present results. As a first study of AI-written journalistic articles in German, it was also shown that GPT-3 texts also perform well in German, even though the language share in the training dataset is only 1.6% (Brown, 2020). Other existing studies found that poems and literature written by the predecessor GPT-2 could also not reliably be distinguished from human originals. Nevertheless, readers generally preferred the human versions over the AI-versions, especially regarding readability and language quality (Gunser et al., 2022; Köbis & Mossink, 2021). This preference for human texts was not the case for the journalistic use case tested in the present study. On the one hand, this implies that AI-written text from the genre of journalistic articles are likely to be better received than from the genre of poetry. On the other hand, as expected, this reflects the better performance of GPT-3 compared to its predecessor GPT-2.

When comparing the results of previous studies involving template-based NLG applications, it becomes evident that the output of these more basic text generators perform a little worse than the AI-based models in terms of linguistic capabilities. In contrast to the AI-written articles tested here, template-based news articles were often perceived as linguistically less proficient and readable than corresponding human texts (Clerwall, 2014; Graefe et al., 2018; Jia, 2020; Graefe & Bohlken, 2020). At this point, the superior linguistic and stylistic possibilities of free text generators like GPT-3 compared to the pre-defined text pieces in the template-based approach, become apparent (see Chapter 2). In terms of perceived credibility and journalistic expertise, template-based texts mostly scored the same or even slightly better than human texts (Graefe et al., 2018; Graefe & Bohlken, 2020; Jia, 2020; Wu, 2019). This was also confirmed for AI-generated articles in the present work. Moreover, GPT-3 was able to keep up with human authors in terms of text quality, although the task of writing a journalistic commentary with its own argumentation was much more demanding than the mostly fact-based use cases of the template-based approaches. The study shows that complex and free argumentations of GPT-3 are perceived by readers as equal to human work, whereas previous research showed this only for simple and fact-delivering reports. However, given these differences in skills and areas of application (see Chapter 2), a comparison of the two NLG approaches is only meaningful to a limited extent. Nevertheless, it reflects the enormous

capabilities of GPT-3 and the progress in the field of AI-generated speech compared to language models used in practice so far.

Although the previous findings regarding the influence of the declared source are not entirely unambiguous, a clear skepticism towards allegedly computer-written texts was nevertheless found. If readers knew about the alleged computer authorship of texts, they were often evaluated more negatively (Graefe & Bohlken, 2020; Longoni et al., 2022; Tewari et al., 2021; Waddell, 2018; Wu, 2019). Readers have thus underestimated the competences of language models (Clark et al., 2021, p. 9). This frequently noted skepticism about the capabilities and trustworthiness of computer-generated content is hardly confirmed in the present work. The disclosure of the alleged text source 'AI' to the readers did not result in lower quality ratings. Only for the evaluation criterion readability the source indication 'human' showed a minimal, non-significant advantage (see Figures 3 & 5). All the other criteria expertise, credibility (intuitive trust) and instrumental trust, were rated about the same for both source declarations, with differences also being insignificant. The previous assumption that the attribution of an article to the source AI, in contrast to a human author, is partly associated with a poorer perception of quality is therefore not supported. This would also mitigate the ethical concerns that media organisations might intentionally omit the true source of computer-written texts in the future (Graefe et al., 2018, p. 605; Graefe & Bohlken, 2020, pp. 57-58). Especially under the assumption that language models will become better and better in the future and become established in society, this issue may not arise.

However, this would have to be tested again with a representative sample of news consumers. The positive perception of allegedly computer-written articles is certainly also related to the composition of the sample after all, which on average shows an almost exactly neutral attitude towards the topic of AI ( $M = 2.9$ ,  $SD = .6$ ). On average, participants were neither particularly skeptical nor over-positive. In combination with the high quality of the AI texts, this resulted in a rating of the texts that was on a par with the human articles. A representative sample would have yielded more reliable results. The extent to which the perception of texts differs among people who are predominantly critical of the topic of AI should be the subject of future research.

The slight positive correlation found between personal attitudes towards AI and the rating of allegedly AI-generated texts (RQ3) suggests that critical people also perceive corresponding texts more negatively and vice versa. The relationship between attitudes towards AI and the rating of the articles was examined. This showed that there is a correlation between the expressed attitudes towards AI and the ratings of the articles that are issued as AI-written. The positive correlations are small to medium, but in comparison with the correlations between the attitudes and the ratings of the texts declared as human, they turn out to be almost three times as large. The strongest

correlation was found with the evaluation criteria credibility ('intuitive trust') with  $r(190) = .31, p < .01$ , as well as instrumental trust  $r(190) = .32, p < .01$  (see Table 8). This is plausible, since the attitudes basically aim at trust towards AI, and these attitudes are then most strongly reflected in the trust placed in the respective texts. Thus, the positive correlation mean that AI-skeptical opinions (regarding its skills and trustworthiness) are to some extent associated with negative evaluations of allegedly AI-produced text. Reversely, optimistic attitudes are to some degree associated with positive evaluations of the corresponding texts.

There is reason to believe that personal attitudes towards AI can have a significant influence on the perception of AI-written texts. However, the present results only show a correlation, whereby no conclusions can be drawn about the causal direction or background of the effect. For example, it remains unclear whether interaction with and reception of AI-produced content shape the corresponding attitudes or vice versa. It can also be assumed that attitudes towards AI are composed of an interplay of diverse experiences. Future research should investigate what other influencing factors exist and what role, for example, concrete prior experience with AI applications, representations in the media, personality traits and one's own prior knowledge or understanding play in this.

## 8.2 Quality criteria and limitations

Essentially, the chosen method of an online experiment in a within-subjects design was suitable for investigating the research interest and testing the hypothesis. Since all participants rated two texts, it was possible to collect sufficient data on the perception of the different conditions and article versions. Based on this data, the descriptive results could be analysed graphically in order to get an overview of the extent to which the article evaluations differed. Beyond that, a multivariate ANOVA was used to test whether the variation of the independent variables real source and declared source of the articles were actually responsible for possible differences. The experimental approach was therefore suitable for testing the concrete causal hypotheses. The fulfilment of the quality criteria objectivity, reliability and internal as well as external validity will be discussed below.

The objectivity of an experiment requires a neutral execution without influencing the test persons and a high degree of standardization (Kubbe, 2016, p. 72). Because the experiment was conducted in an online setting, only minimal intervention by the researcher was necessary. There was no direct, personal interaction with the test persons and no opinion or preconception was expressed. Only the participation request and the necessary instructions in the experimental procedure were given

in neutral language. Furthermore, a standard procedure was chosen for the experiment, based on a previously successful study (Graefe et al., 2018). The statistical evaluation to test the hypothesis was also conducted by a standardized and objective manner. Objectivity of the results is thus given.

How reliable an experiment is depends on how replicable the results are under the same conditions (Kubbe, 2016, p. 71). The clear description of the experimental design and procedure (Bryman, 2012, p. 54) makes replication generally possible. In principle, it would be expected that subjects would perhaps perceive the articles differently if the experiment was carried out again, and that their attitudes towards AI might change. The results are a momentary observation. Both readers' expectations and attitudes towards AI-written content as well as AI-applications themselves evolve fast. This will possibly make for a better perceived text quality in the future (Graefe & Bohlken, 2020, p. 58). The ratings may of course also differ for other topics or contents than those used here. The study shows a high reliability for the experimental design. With regard to the survey of the perception of the articles, all four measurement constructs for article quality used turned out to be reliable with values of Cronbach's alpha of around  $\alpha = .8$  (see Chapter 6.8). For the measurement of the respondents' attitudes towards AI, the reliability value for the scale used was only moderate ( $\alpha = .61$ ). Therefore, the findings regarding the influence of attitudes on the evaluation of the texts (RQ3) can only be interpreted to a limited extent. In view of this, future research should update the item set for increased reliability.

Of particular relevance in experimental studies is the validity, which refers to whether the varied dependent variable (real and declared source) can be said to be exclusively responsible and causal for differing measurement values in the dependent variable (perception of the articles) (Dickmann, 2005, p. 297; Kubbe, 2016, p. 61). The study features a fairly high internal validity, which means that the causal explanations for the case at hand are quite trustworthy. A true experiment (Bryman, 2012, p. 50) was conducted as the design included random group assignment, a comparison group for each manipulated variable (declared and real source) and the independent variable preceded the dependent variable in the experimental setup (Bryman, 2012, p. 53; Sedlmeier & Renkewitz, 2018, p. 133). All necessary conditions of an experimental design were thus met. Additionally, it was checked whether there were systematically accumulated dropouts in individual conditions, which can compromise the correct randomisation and the neutralisation of person-related confounding factors. This was not the case (see Chapter 6.8). It can therefore be assumed that all experimental conditions were undergone by groups with comparable backgrounds and personal characteristics. In addition, person-related confounding variables were neutralised anyway by the within-subjects design (Sedlmeier & Renkewitz, 2018, p. 157). For the independent variable 'declared source', all

subjects went through both possible variations ‘AI’ and ‘human’ (see Table 4). Confounding variables were therefore sufficiently controlled and a possible influence on the dependent variables (article rating) could be avoided. Possible differences in the item ratings can thus be attributed to the variation of the declared and real source.

Internal validity, in other words the correct measuring of what one wants to measure, also depends on whether the experimental manipulation of the independent variable worked at all and was perceived (Bryman, 2012, p. 53). In this respect, no explicit manipulation check was carried out in the questionnaire, as for example in the work of Waddell (2018). In order to keep the questionnaire as simple and short as possible, no control question was asked as to whether respondents actively perceived the declared source as stated. Nevertheless, it can be assumed that the manipulation was successful, since according to the experimental instructions, the subjects believed they were making an evaluation of a human text on the one hand and an AI-written text on the other (see Chapter 6.2.2). It was thus made clear to the subjects that one of the texts was supposedly AI-written and the other was not. Accordingly, two opposing states (declared source AI or human), are already present in maximum expression. A further, gradual increase of the independent variables is not possible. Since the expressions are therefore in natural opposition to each other, it must be assumed that the manipulation is successful (Huber et al., 2014, p. 62).

Since an online experiment takes place in a less well-controlled environment compared to the laboratory setting, technical disturbances or other distractions of all kinds cannot be ruled out (Gnambs & Strassnig, 2007, pp. 237-238), which may affect internal validity negatively. In addition, people may not participate attentively or seriously enough (Gnambs & Strassnig, 2007, p. 239). Even though such situationally specific confounding variables (Gnambs & Strassnig, 2007, p. 233) should be evenly distributed across all experimental conditions through randomisation and thus controlled, a small residual risk remains. Internal validity was thus ensured as best as possible. However, one can never be completely sure of absolute internal validity, as most confounding variables are unknown (Sedlmeier & Renkewitz, 2018, p. 138).

However, conducting the experiment in an online environment has a positive effect on external validity. The designed experimental situation was close to a real-life use case of reading news. The study was conducted on the participants’ own smartphone, tablet or laptop in a private environment. Actual news reception often takes place under very similar conditions. The real situation of news consumption was therefore simulated as best as possible in the experiment, which means that external validity is higher than in a laboratory experiment, for example. The results are therefore particularly valid for the case of digitally consumed articles. The absence of a supervisor

in the online experiment is not only beneficial for objectivity but also for external validity. Subjects complete the test in the online experimental setting a more familiar environment, compared to laboratory experiments, and are as close as possible to real conditions (Gnambs & Strassnig, 2007, p. 237).

It is important to note that test subjects are generally aware that they are taking part in an experiment, which can influence their response behaviour. This has a negative effect on the transferability of results from experimental studies to larger populations and thus on external or ecological validity (Bryman, 2012, p. 54). Furthermore, there is no representative sample for the population of German-speaking readers, which limits external validity. The sample is comparatively young, educated above average and, according to their own statements, relatively well informed about AI (see Figure 1). In addition, the attitudes towards AI are on average fairly neutral ( $M = 2.9$ ,  $SD = .6$ ). AI-critical people might have rated the AI texts more negatively. The results are therefore particularly well transferable to an educated and young group without major resentment towards the topic of Artificial Intelligence (AI).

However, while appreciating the positive performance of the AI texts, it should be noted that a high level of human involvement was necessary in the article generation process. The articles tested in the experiment were versions reviewed and curated by the researcher and not just any result produced by GPT-3. The human involvement in the writing process consisted, on the one hand, of giving the AI the opinion or basic direction of argumentation. This was done through the beginning of the human article, which was given as input. On the other hand, the language model is not bound to truth which resulted in regularly committed errors of grammatical or factual nature while generating the text samples. The model also completely makes up aspects sometimes. Where necessary, the language model was given several chances to correct possible errors in the text by deleting the erroneous passage and generating text again from there. In this way, a quality assurance was carried out, without which such convincing articles would not have been possible (for more details see Chapter 6.5). In a way, the qualitative optimum of output was tested. Since the AI as a black box delivers unpredictable results (Giansiracusa, 2021, p. 33), an appropriate selection of the output is essential, as previous work has also shown (Köbis & Mossink, 2021, p. 10). GPT-3 is an intelligent model that speaks human languages remarkably well, but whose contents cannot easily be trusted (Marcus & Davis, 2020).

As long as the error-proneness and unpredictability of AI language output remains, humans will be necessary in the writing and publishing process. Since GPT-3 may invent facts and even

appropriate sources, well-informed people are still needed as fact-checkers. In the future, hybrid forms of collaboration may be possible, where language models generate lots of content and different versions of articles that are edited by humans and selected for publication (Köbis & Mossink, 2021, p. 11). This could lead to a paradigm shift in journalism and the media industry, where staff need to do less texting themselves but need more extensive knowledge on how to use such language systems (Graefe, 2016).

Apart from concrete applications in media and communication and in order to develop independent and universal artificial intelligence, it is therefore necessary to continue to work on a real understanding of the world. To achieve this, the ability of logic reasoning, causal understanding and knowledge of facts needs to be implemented. The current purely statistical language completion of GPT-3 is only one part among many of an actual artificial intelligence (Elkins & Chun, 2020, p. 13).

Despite the mentioned limitations, the results show the enormous potential of such self-learning language models, which will continue to advance, especially through rapid development in AI. With ‘chatGPT’, the chatbot version of GPT-3 released while this study was conducted and its successor GPT-4 just around the corner, the rapid development of language AIs that we have seen so far will continue. AI-produced content will become increasingly popular in the media and online. Whether and how quickly readers will get used to this new content source in practice remains to be investigated in the future. The results of the present work also show that GPT-3 produced content that is at least reviewed and worthy of publication can hardly be distinguished from human content any more. What influence these new circumstances will have on the relevance of the origin of content at all in the future will have to be clarified (Floridi & Chiriatti, 2020, p. 691). After the emergence of template-based systems for automating basic texts in the 2010s (Graefe, 2016, pp. 21-31), AI-based language models now mark the next step towards sophisticated language production and the merging of humans and machines (Graefe, 2016, p. 11) in content production.

## 9 Conclusion

The study investigated how journalistic articles written by the self-learning language model GPT-3 are perceived compared to human-written versions. From two opinion pieces by human journalists on two different topics, an AI-written version of the two articles was created using GPT-3. On that basis, an experiment in a 2x2x2 within-subjects design was conducted, varying the topics, as well as the actual and the declared source (human-written or AI-written) of articles. Participants rated the articles regarding the quality criteria ‘readability’, ‘journalistic expertise’, ‘credibility’ (‘intuitive trust’) and ‘instrumental trust’.

The results demonstrated that journalistic opinion pieces written in German by GPT-3 can hardly be distinguished by readers from the corresponding human versions. For one, the study found that readers considered articles written by the artificial intelligence GPT-3 to be just as good as those written by human journalists. The AI-generated versions were even rated slightly better in terms of all rating criteria of readability, expertise, credibility, and trust. The text quality as well as the trust placed in the content and the argumentation were thus perceived as qualitatively equal. However, differences were not statistically significant. Varying the actual source thus had no effect on the general article perception. For another, the idea that readers would be skeptical of articles written by AI was not supported by the results. Fundamentally, the study participants showed a neutral attitude towards the topic of AI and its capabilities. Barely any negative effect on the rating of texts could be found if the readers assumed that they were reading a text produced by an AI. Articles declared as AI-written were mostly rated equally well or even minimally better than texts declared as human (for expertise credibility, trust). Only the readability was rated slightly worse if the AI was stated as a source of the articles instead of a human. Accordingly, the variation of the declared source had only a small negative effect on readability for the case of AI-declared texts. For one, it can be concluded that AI-written texts were well accepted on average in the sample. For another, it remains questionable how a generally more AI-skeptical group would have reacted. Moreover, a relationship was found between the participants’ personal attitudes towards AI and their perception of articles allegedly written by an AI. For articles declared as AI-written, the personal attitudes correlate slightly to moderately with all of the four rating criteria. In contrast, no significant correlation was found for the allegedly human-written articles. This suggests that AI-skeptical opinions (regarding its trustworthiness and skills) are to some extent associated with negative evaluations of allegedly AI-produced text. Vice versa, optimistic attitudes are to some degree associated with positive evaluations of the corresponding texts. This suggests that personal attitudes towards the topic of AI may be an influential factor on how AI-produced content is



perceived. This raises the question as to what other influencing factors exist and what role concrete prior experiences with AI applications or personal knowledge on the subject of AI play. In the light of the intensive reporting on the AI assistant chatGPT, which is based on GPT-3 and was published in late 2022, the question of how media reporting on the topic influences reader attitudes towards AI-produced content is becoming increasingly important.

In this study, the AI language model GPT-3 was able to produce articles that were perceived equal to human versions in terms of all evaluation criteria. This implies that readers were hardly able to distinguish between AI-written and human articles, which is in line with previous findings on GPT-3-produced articles by Brown et al. (2020) and Clark et al. (2021). While similar effects were already shown in previous studies for automated texts in mostly fact-based use cases, the present work focused on articles including explicit opinion statements and arguments concerning the respective topics. For the first time, it was investigated how an AI language model performs on the task of writing a journalistic commentary on a topic, including its own arguments and logic reasoning. Despite the demanding task of writing an opinion piece, GPT-3 articles could not only keep up with human versions in all respects, but partly even surpassed them insignificantly. GPT-3's writing and reasoning qualities were thus perceived by readers to be as good as those of humans, which underlines the enormous potential of AI-powered NLG solutions.

Nevertheless, at this stage, there is no way that such language models could take over the entire work of human journalists. The most relevant limitation of the results is that a quite high level of human involvement was needed to generate the articles. GPT-3 is not committed to truth and may just invent facts. It regularly committed grammatical or logical errors when producing the tested articles. The language model needed various attempts and was given several opportunities to correct itself in order to produce the final text. Articles were thus selected based on human judgement. Without close human supervision, such a good output would not have been possible. In addition, the basic argumentative approach was partly already provided within the initial input to the model by the human text. Humans are therefore still essential as fact-checkers and for quality assessment, above all in journalism (Dale, 2021, p. 116; Elkins & Chun, 2020, p. 8). Collaborating with language models and learning how to apply them for specific use cases will become more important in journalism and the media industry. Language models do not yet enable complete automation of writing, but rather represent a partial automation in which the first drafts are computer-generated and then edited. Already today, a new discipline known as prompt engineering is slowly emerging, in which the task is to formulate the instructions (prompt) and the initial input for GPT-3 in such a way that the desired output is achieved (Shieh, n.d.).

To conclude, it can be argued that GPT-3 is capable of speaking human languages remarkably well, but the correctness of the output cannot be easily trusted. It should at this point be considered a mere language model and not a general artificial intelligence. Considering its linguistically compelling results shown in this study on the one hand as well as its error-proneness on the other hand, one of the biggest issues of GPT-3 turns out to be reliability. The linguistic qualities of self-learning language models like GPT3 are therefore only one component on the way to general AI (Elkins & Chun, 2020, p. 13). For a more trustworthy intelligence, however, other skills such as actual fact-based knowledge are still required. In order to develop independent and universal artificial intelligence, it is therefore necessary to continue to work on a real understanding of the world. To achieve this, the ability of reliable logic reasoning, causal understanding and knowledge of facts needs to be developed. The current purely statistical language completion of GPT-3 is not sufficient to create a real artificial intelligence.

To evaluate future developments in this direction, research on the linguistic and content quality of output of AI language models in different use cases will remain essential. This also raises the question of user perception of future machine-generated content, how attitudes towards machine-generated content shift and what influence the increasing presence of computer-generated content will have on media landscapes and societies.

## Bibliography

- Abid, A., Farooqi, M., & Zou, J. (2021). Persistent Anti-Muslim Bias in Large Language Models. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 298–306.  
<https://doi.org/10.1145/3461702.3462624>
- Bitkom Research. (2017, November 15). *Künstliche Intelligenz*. bitkom - Künstliche Intelligenz, Berlin. <https://www.bitkom.org/sites/default/files/file/import/Bitkom-Charts-PK-AI-15-11-2017-final.pdf>
- Bortz, J. (2005). *Statistik für Human- und Sozialwissenschaftler: Mit 242 Tabellen* (6.). Springer Medizin.
- Brown, T. B. (2020). *GPT-3: Public archive*. OpenAI. [https://github.com/openai/gpt-3/blob/d7a9bb505df6f630f9bab3b30c889e52f22eb9ea/dataset\\_statistics/languages\\_by\\_character\\_count.csv](https://github.com/openai/gpt-3/blob/d7a9bb505df6f630f9bab3b30c889e52f22eb9ea/dataset_statistics/languages_by_character_count.csv) (Original work published 2020)
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). *Language Models are Few-Shot Learners* (arXiv:2005.14165). arXiv.  
<http://arxiv.org/abs/2005.14165>
- Bryman, A. (2012). *Social research methods* (4th ed). Oxford University Press.
- Brybaert, M. (2019). How Many Participants Do We Have to Include in Properly Powered Experiments? A Tutorial of Power Analysis with Reference Tables. *Journal of Cognition*, 2(1): 16), 1–38. <https://doi.org/10.5334/joc.72>
- Chien, S.-Y., Lewis, M., Semnani-Azad, Z., & Sycara, K. (2014). An Empirical Model of Cultural Factors on Trust in Automation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 58(1), 859–863. <https://doi.org/10.1177/1541931214581181>
- Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., & Smith, N. A. (2021). *All That's Human' Is Not Gold: Evaluating Human Evaluation of Generated Text* (arXiv:2107.00061). arXiv. <http://arxiv.org/abs/2107.00061>

- Clerwall, C. (2014). Enter the Robot Journalist: Users' perceptions of automated content. *Journalism Practice*, 8(5), 519–531. <https://doi.org/10.1080/17512786.2014.883116>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). L. Erlbaum Associates.
- Dale, R. (2020). Natural language generation: The commercial state of the art in 2020. *Natural Language Engineering*, 26(4), 481–487. <https://doi.org/10.1017/S135132492000025X>
- Dale, R. (2021). GPT-3: What's it good for? *Natural Language Engineering*, 27(1), 113–118. <https://doi.org/10.1017/S1351324920000601>
- Diekmann, A. (2005). *Empirische Sozialforschung: Grundlagen, Methoden, Anwendungen* (17. Auflage). Rowohlt Taschenbuch Verlag.
- Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humannwissenschaften*. Springer. <https://doi.org/10.1007/978-3-642-41089-5>
- Dou, Y., Forbes, M., Koncel-Kedziorski, R., Smith, N., & Choi, Y. (2022). Is GPT-3 Text Indistinguishable from Human Text? Scarecrow: A Framework for Scrutinizing Machine Text. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7250–7274. <https://doi.org/10.18653/v1/2022.acl-long.501>
- Dzcyk, W. (2001). Ethische Dimensionen der Online-Forschung. *Kölner psychologische Studien*, 6(1).
- Eliot, D. L. B. (2021). *AI & Law | The Master Class Series*. 7.
- Elkins, K., & Chun, J. (2020). Can GPT-3 Pass a Writer's Turing Test? *Journal of Cultural Analytics*, 5(2). <https://doi.org/10.22148/001c.17212>
- Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge University Press.
- Floridi, L., & Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines*, 30(4), 681–694. <https://doi.org/10.1007/s11023-020-09548-1>
- Garms-Homolová, V. (2020). *Sozialpsychologie der Einstellungen und Urteilsbildung: Lässt sich menschliches Verhalten vorhersagen?* Springer. <https://doi.org/10.1007/978-3-662-62434-0>

- Gatt, A., & Krahmer, E. (2018). Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61, 65–170. <https://doi.org/10.1613/jair.5477>
- Gharat, A., Tandel, H., & Bagade, K. (2019). Natural Language Processing Theory, Applications and Difficulties. *International Journal of Trend in Scientific Research and Development (IJTSRD)*, 3(6), 501–503.
- Giansiracusa, N. (2021). *How Algorithms Create and Prevent Fake News: Exploring the Impacts of Social Media, Deepfakes, GPT-3, and More*. Apress. <https://doi.org/10.1007/978-1-4842-7155-1>
- Glahn, H. R. (1970). Computer-produced worded forecasts. *Bulletin American Meteorological Society*, 51(12), 1126–1132.
- Gnambs, T., & Strassnig, B. (2007). Experimentelle Online-Untersuchungen. In M. Welker, O. Wenzel, & Deutsche Gesellschaft für Online-Forschung (Eds.), *Online-Forschung 2007: Grundlagen und Fallstudien* (pp. 232–249). Herbert von Halem.
- Graefe, A. (2016). *Guide to Automated Journalism*.
- Graefe, A., & Bohlken, N. (2020). Automated Journalism: A Meta-Analysis of Readers' Perceptions of Human-Written in Comparison to Automated News. *Media and Communication*, 8(3), 50–59. <https://doi.org/10.17645/mac.v8i3.3019>
- Gunser, V. E., Gottschling, S., Brucker, B., Richter, S., Çakir, D. C., & Gerjets, P. (2022). The Pure Poet: How Good is the Subjective Credibility and Stylistic Quality of Literary Short Texts Written with an Artificial Intelligence Tool as Compared to Texts Written by Human Authors? *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*, 60–61. <https://doi.org/10.18653/v1/2022.in2writing-1.8>
- Hauptmanns, P. (1999). Grenzen und Chancen von quantitativen Befragungen mit Hilfe des Internet. In B. Batinic, A. Werner, L. Gräf, & W. Bandilla (Eds.), *Online-research: Methoden, Anwendungen und Ergebnisse* (pp. 21–38). Hogrefe.

- Ho, G. (2019, March 9). Autoregressive Models in Deep Learning—A Brief Survey. *George Ho*.  
<https://www.georgeho.org/deep-autoregressive-models/>
- Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). *The Curious Case of Neural Text Degeneration* (arXiv:1904.09751). arXiv. <http://arxiv.org/abs/1904.09751>
- Huber, F., Meyer, F., & Lenzen, M. (2014). Auswertung von Experimenten mittels Varianzanalyse. In F. Huber, F. Meyer, & M. Lenzen, *Grundlagen der Varianzanalyse* (pp. 43–88). Springer Fachmedien. [https://doi.org/10.1007/978-3-658-05666-7\\_4](https://doi.org/10.1007/978-3-658-05666-7_4)
- Jia, C. (2020). Chinese Automated Journalism: A Comparison Between Expectations and Perceived Quality. *International Journal of Communication*, 14(2020), 2611–2632.
- Kilg, M., Meier, P., Graw, L., Salzer, R., & Körner, T. (2022, November 26). Kolumne einer künstlichen Intelligenz: Wie schreibt eine Robo-Autor\*in? *taz, die tageszeitung*.  
<https://taz.de/15898282/>
- Köbis, N., & Mossink, L. D. (2021). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in Human Behavior*, 114, 106553. <https://doi.org/10.1016/j.chb.2020.106553>
- Kubbe, I. (2016). *Experimente in der Politikwissenschaft*. Springer Fachmedien.  
<https://doi.org/10.1007/978-3-658-09424-9>
- Latar, N. L. (2015). The Robot Journalist in the Age of Social Physics: The End of Human Journalism? In G. Einav (Ed.), *The New World of Transitioned Media* (pp. 65–80). Springer International Publishing. [https://doi.org/10.1007/978-3-319-09009-2\\_6](https://doi.org/10.1007/978-3-319-09009-2_6)
- Longoni, C., Fradkin, A., Cian, L., & Pennycook, G. (2022). News from Generative Artificial Intelligence Is Believed Less. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 97–106. <https://doi.org/10.1145/3531146.3533077>
- Lucy, L., & Bamman, D. (2021). Gender and Representation Bias in GPT-3 Generated Stories. *Proceedings of the Third Workshop on Narrative Understanding*, 48–55.  
<https://doi.org/10.18653/v1/2021.nuse-1.5>

- Mann, B. (2019, May 30). How to sample from language models. *Towards Data Science*.  
<https://towardsdatascience.com/how-to-sample-from-language-models-682bceb97277>
- Marcus, G., & Davis, E. (2020, August 22). *GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about*. <https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/>
- McGuffie, K., & Newhouse, A. (2020). *The Radicalization Risks of GPT-3 and Advanced Neural Language Models* (arXiv:2009.06807). arXiv. <http://arxiv.org/abs/2009.06807>
- Menold, N., & Bogner, K. (2015). *Gestaltung von Ratingskalen in Fragebögen (Version 1.1)*. (GESIS *Survey Guidelines*). [https://doi.org/10.15465/GESIS-SG\\_015](https://doi.org/10.15465/GESIS-SG_015)
- Montal, T., & Reich, Z. (2017). I, Robot. You, Journalist. Who is the Author?: Authorship, bylines and full disclosure in automated journalism. *Digital Journalism*, 5(7), 829–849.  
<https://doi.org/10.1080/21670811.2016.1209083>
- O'Muircheartaigh, C., Krosnick, J. A., & Helic, A. (2000). *Middle Alternatives, Acquiescence, and the Quality of Questionnaire Data*. Annual Meeting of the American Association for Public Opinion Research, St. Petersburg, Florida.
- OpenAI. (n.d.-a). *Completions*. OpenAI - Documentation. Retrieved 2 February 2023, from <https://platform.openai.com/docs/api-reference/completions/create>
- OpenAI. (n.d.-b). *Introduction*. OpenAI - Documentation. Retrieved 5 February 2023, from <https://platform.openai.com/docs/introduction/overview>
- OpenAI. (n.d.-c). *Limitations & risks*. OpenAI - Documentation. Retrieved 2 February 2023, from <https://platform.openai.com/docs/guides/embeddings/limitations-risks>
- OpenAI. (n.d.-d). *Models*. OpenAI - Documentation. Retrieved 2 February 2023, from <https://platform.openai.com/docs/models>
- OpenAI. (n.d.-e). *OpenAI API (Playground)*. Retrieved 5 February 2023, from <https://platform.openai.com/playground>

- OpenAI. (n.d.-f). *Parameter details*. OpenAI - Documentation. Retrieved 2 February 2023, from <https://platform.openai.com/docs/api-reference/parameter-details>
- OpenAI. (n.d.-g). *Quickstart*. OpenAI - Documentation. Retrieved 2 February 2023, from <https://platform.openai.com/docs/quickstart>
- OpenAI. (2020, June 11). OpenAI API. *OpenAI*. <https://openai.com/blog/openai-api/>
- OpenAI. (2022, November 30). ChatGPT: Optimizing Language Models for Dialogue. *OpenAI*. <https://openai.com/blog/chatgpt/>
- Perera, R., & Nand, P. (2017). Recent Advances in Natural Language Generation: A Survey and Classification of the Empirical Literature. *Computing and Informatics*, 36(1), 1–32. [https://doi.org/10.4149/cai\\_2017\\_1\\_1](https://doi.org/10.4149/cai_2017_1_1)
- Perrotta, C., Selwyn, N., & Ewin, C. (2022). Artificial intelligence and the affective labour of understanding: The intimate moderation of a language model. *New Media & Society*, 0(0). <https://doi.org/10.1177/14614448221075296>
- Porst, R. (2014). *Fragebogen*. Springer VS. <https://doi.org/10.1007/978-3-658-02118-4>
- Prochazka, F. (2020). *Vertrauen in Journalismus unter Online-Bedingungen: Zum Einfluss von Personenmerkmalen, Qualitätswahrnehmungen und Nachrichtennutzung*. Springer Fachmedien. <https://doi.org/10.1007/978-3-658-30227-6>
- Recchia, G. (2021). Teaching Autoregressive Language Models Complex Tasks By Demonstration. *ArXiv*, 15. <https://doi.org/10.48550/ARXIV.2109.02102>
- Reips, U.-D. (2002). Standards for Internet-Based Experimenting. *Experimental Psychology*, 49(4), 243–256. <https://doi.org/10.1027//1618-3169.49.4.243>
- Reips, U.-D. (2021). Web-Based Research in Psychology: A Review. *Zeitschrift Für Psychologie*, 229(4), 198–213. <https://doi.org/10.1027/2151-2604/a000475>
- Reiter, E., & Dale, R. (1997). Building applied natural language generation systems. *Natural Language Engineering*, 3(1), 57–87. <https://doi.org/10.1017/S1351324997001502>



- Rohrman, B. (2007). *Verbal qualifiers for rating scales: Sociolinguistic considerations and psychometric data*. Project Report. University of Melbourne.  
<http://www.rohrmannresearch.net/pdfs/rohrmannvqs-report.pdf>
- Sedlmeier, P., & Renkewitz, F. (2018). *Forschungsmethoden und Statistik für Psychologen und Sozialwissenschaftler* (3. Auflage). Pearson.
- Shieh, J. (n.d.). Best practices for prompt engineering with OpenAI API. *Advice and Answers from the OpenAI Team*. Retrieved 5 February 2023, from  
<https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-openai-api>
- Sobieszek, A., & Price, T. (2022). Playing Games with Ais: The Limits of GPT-3 and Similar Large Language Models. *Minds and Machines*, 32(2), 341–364.  
<https://doi.org/10.1007/s11023-022-09602-0>
- SoSciSurvey. (2019, June 4). *Randomisierung mit Urnen*. SoSciSurvey.  
[https://www.soscisurvey.de/help/doku.php/de:create:random\\_urns](https://www.soscisurvey.de/help/doku.php/de:create:random_urns)
- Sundar, S. S. (1999). Exploring Receivers' Criteria for Perception of Print and Online News. *Journalism & Mass Communication Quarterly*, 76(2), 373–386.  
<https://doi.org/10.1177/107769909907600213>
- Tewari, S., Zabounidis, R., Kothari, A., Bailey, R., & Alm, C. O. (2021). Perceptions of Human and Machine-Generated Articles. *Digital Threats: Research and Practice*, 2(2), 1–16.  
<https://doi.org/10.1145/3428158>
- Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response* (1st ed.). Cambridge University Press. <https://doi.org/10.1017/CBO9780511819322>
- Tschopp, M., Ruef, M., & Monett, D. (2022). Vertrauen Sie KI? Einblicke in das Thema Künstliche Intelligenz und warum Vertrauen eine Schlüsselrolle im Umgang mit neuen Technologien spielt. In M. Landes, E. Steiner, & T. Utz (Eds.), *Kreativität und Innovation in*

- Organisationen* (pp. 319–346). Springer Berlin Heidelberg. [https://doi.org/10.1007/978-3-662-63117-1\\_16](https://doi.org/10.1007/978-3-662-63117-1_16)
- van der Kaa, H. A. J., & Kraemer, E. J. (2014). Journalist versus news consumer: The perceived credibility of machine written news. *Proceedings of the Computation+Journalism Conference*. <https://research.tilburguniversity.edu/en/publications/journalist-versus-news-consumer-the-perceived-credibility-of-mach>
- van der Lee, C., Gatt, A., van Miltenburg, E., Wubben, S., & Kraemer, E. (2019). Best practices for the human evaluation of automatically generated text. *Proceedings of the 12th International Conference on Natural Language Generation*, 355–368. <https://doi.org/10.18653/v1/W19-8643>
- Waddell, T. F. (2018). A Robot Wrote This?: How perceived machine authorship affects news credibility. *Digital Journalism*, 6(2), 236–255. <https://doi.org/10.1080/21670811.2017.1384319>
- Wilson, P. (1983). *Second-hand knowledge: An inquiry into cognitive authority*. Greenwood Press.
- Wu, Y. (2019). Is Automated Journalistic Writing Less Biased? An Experimental Test of Auto-Written and Human-Written News Stories. *Journalism Practice*, 14(8), 1008–1028. <https://doi.org/10.1080/17512786.2019.1682940>
- Wullenkord, R. (2017). *Messung und Veränderung von Einstellungen gegenüber Robotern – Untersuchung des Einflusses von imaginiertem Kontakt auf implizite und explizite Maße*. Universität Bielefeld.
- Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2020). Defending Against Neural Fake News. *ArXiv:1905.12616 [Cs]*. <http://arxiv.org/abs/1905.12616>
- Zydorek, C. (2018). *Grundlagen der Medienwirtschaft*. Springer Fachmedien. <https://doi.org/10.1007/978-3-658-15252-9>

# Appendix

## List of appendices

Appendix 1: Mean article ratings across all experimental conditions.....	84
Appendix 2: MANOVA results.....	85
Appendix 3: Experimental treatments (journalistic articles) .....	86
Appendix 3.1: Human-written article topic 1 (Olympics).....	86
Appendix 3.2: AI-written article topic 1 (Olympics).....	88
Appendix 3.3: Human-written article topic 2 (Queen).....	90
Appendix 3.4: AI-written article topic 2 (Queen).....	92
Appendix 4: Exemplary experimental procedure.....	94
Appendix 4.1: Introduction to the survey .....	94
Appendix 4.2: Survey of personal attitudes towards the topic of AI.....	95
Appendix 4.3: Explanation regarding the first article.....	96
Appendix 4.4: Presentation of the first article .....	96
Appendix 4.5: Evaluation of the first article .....	97
Appendix 4.6: Explanation regarding the second article.....	98
Appendix 4.7: Presentation of the first article .....	98
Appendix 4.8: Evaluation of the second article .....	99
Appendix 4.9: Querying demographic data.....	100
Appendix 4.10: Debriefing .....	101

## Appendix 1: Mean article ratings across all experimental conditions

Topic	Real source	Declared source	<i>N</i>	readability	expertise	credibility	instrumental trust
Olympics	Human	Human	49	3.2 (0.86)	3.2 (0.79)	3.4 (0.76)	2.8 (0.75)
		Artificial Intelligence	51	3.1 (0.91)	3.3 (0.83)	3.2 (0.80)	2.9 (0.80)
		Total	100	3.2 (0.89)	3.3 (0.81)	3.3 (0.78)	2.9 (0.77)
	Artificial Intelligence	Human	45	3.4 (0.78)	3.7 (0.81)	3.5 (0.67)	3.1 (0.81)
		Artificial Intelligence	47	3.2 (0.91)	3.7 (0.78)	3.6 (0.74)	3.2 (0.80)
		Total	92	3.3 (0.85)	3.7 (0.79)	3.6 (0.70)	3.1 (0.80)
	Total	Human	94	3.3 (0.82)	3.4 (0.83)	3.4 (0.72)	3.0 (0.79)
		Artificial Intelligence	98	3.2 (0.91)	3.5 (0.82)	3.4 (0.79)	3.0 (0.80)
		Total	192	3.2 (0.87)	3.5 (0.83)	3.4 (0.76)	3.0 (0.80)
Queen	Human	Human	49	3.5 (1.06)	3.4 (0.92)	3.3 (0.80)	3.0 (1.07)
		Artificial Intelligence	49	3.3 (0.97)	3.5 (0.72)	3.5 (0.67)	3.1 (0.87)
		Total	98	3.4 (1.02)	3.5 (0.82)	3.4 (0.74)	3.0 (0.97)
	Artificial Intelligence	Human	49	3.6 (0.77)	3.4 (0.80)	3.4 (0.76)	3.0 (0.93)
		Artificial Intelligence	45	3.4 (0.85)	3.6 (0.71)	3.5 (0.66)	3.0 (0.83)
		Total	94	3.5 (0.81)	3.5 (0.76)	3.4 (0.71)	3.0 (0.88)
	Total	Human	98	3.6 (0.92)	3.4 (0.86)	3.4 (0.78)	3.0 (1.00)
		Artificial Intelligence	94	3.3 (0.91)	3.6 (0.71)	3.5 (0.66)	3.0 (0.85)
		Total	192	3.5 (0.92)	3.5 (0.79)	3.4 (0.72)	3.0 (0.93)
Total	Human	Human	98	3.4 (0.97)	3.3 (0.86)	3.3 (0.78)	2.9 (0.93)
		Artificial Intelligence	100	3.2 (0.95)	3.4 (0.86)	3.3 (0.75)	3.0 (0.83)
		Total	198	3.3 (0.96)	3.4 (0.86)	3.3 (0.76)	3.0 (0.88)
	Artificial Intelligence	Human	94	3.5 (0.78)	3.5 (0.81)	3.4 (0.72)	3.0 (0.88)
		Artificial Intelligence	92	3.3 (0.88)	3.7 (0.74)	3.5 (0.70)	3.1 (0.82)
		Total	186	3.4 (0.83)	3.6 (0.78)	3.5 (0.71)	3.1 (0.84)
	Total	Human	192	3.4 (0.88)	3.4 (0.84)	3.4 (0.75)	3.0 (0.90)
		Artificial Intelligence	192	3.2 (0.91)	3.5 (0.77)	3.4 (0.73)	3.0 (0.82)
		Total	384	3.3 (0.90)	3.5 (0.81)	3.4 (0.74)	3.0 (0.86)

*Note.* Mean ratings based on a 5-point scale (1 = ‘Do not agree at all’, 5= ‘Fully agree’). Standard errors in brackets.

## Appendix 2: MANOVA results

Results from a multi-factorial, multivariate analysis of variance

Independent Variable	Dependent variable	<i>F</i>	Wilk's $\Lambda$	<i>p</i>	$\eta^2_{part}$
Real source		$F(4, 373) = 2.219$	.977	.066	.023
	Readability	$F(1, 376) = 1.501$		.221	.004
	Expertise	$F(1, 376) = 7.578$		.006	.020
	Credibility	$F(1, 376) = 3.406$		.066	.009
	Instrumental trust	$F(1, 376) = 1.470$		.226	.004
Declared source		$F(4, 373) = 3.769$	.961	.005	.039
	Readability	$F(1, 376) = 3.880$		.050	.010
	Expertise	$F(1, 376) = 2.352$		.126	.006
	Credibility	$F(1, 376) = 0.132$		.717	.000
	Instrumental trust	$F(1, 376) = 0.238$		.626	.001
Real x declared source		$F(4, 373) = 0.262$	.997	.902	.003
	Readability	-		-	-
	Expertise	-		-	-
	Credibility	-		-	-
	Instrumental trust	-		-	-
Topic		$F(4, 373) = 2.483$	.974	.043	.026
	Readability	$F(1, 376) = 5.855$		.016	.015
	Expertise	-		-	-
	Credibility	-		-	-
	Instrumental trust	-		-	-

*Note.* Results from a multi-factorial, multivariate analysis of variance (multi-factorial MANOVA), including ratings for both article versions. Results for the four dependent variables are univariate analyses.

## Appendix 3: Experimental treatments (journalistic articles)

### Appendix 3.1: Human-written article topic 1 (Olympics)

#### **Entschädigung nach Olympia-Attentat**

### **Angemessen und überfällig**

Dass die Hinterbliebenen der Opfer des Olympia-Attentats die Gedenkfeier als Druckmittel für ihre Forderungen nutzten, ist umstritten. Doch Deutschland hätte es auch niemals so weit kommen lassen dürfen.

Ein Kommentar von **Hannah Berger**

Der ganz große Eklat wurde vermieden. Dass am 5. September in Deutschland der Opfer des Münchner Olympia-Anschlags gedacht wird und die Familien der ermordeten israelischen Sportler aus Protest mehrheitlich fernbleiben - dieses Szenario haben die Bundesregierung, der Freistaat Bayern und die Stadt München nun also abgewendet. Indem sie dem Vernehmen nach rund 28 Millionen Euro Entschädigung für die Opferangehörigen zusagten.

Den Familien war bewusst, welche Bedeutung die Gedenkveranstaltung für die deutsche Seite hat und sie nutzten diese Machtposition im Streit um Entschädigungszahlungen aus. Ja, dass das Gedenken an die Opfer mit finanziellen Zielen verbunden wird, darf man irritierend finden. Die Vorgehensweise ist auch in Israel nicht unumstritten. Aber hatten die Familien eine andere Wahl?

Das Versagen der deutschen Sicherheitsbehörden und damit die Mitverantwortung für den blutigen Ausgang der Geiselnahme sind unstrittig. Die bisher von deutscher Seite gezahlten Summen waren angesichts von elf betroffenen Familien allerdings lächerlich niedrig und wurden auch immer als Unterstützungszahlungen deklariert. Das Wort Entschädigung wurde stets vermieden, dabei war es doch so angebracht.

Seit Jahrzehnten fordern die Opferangehörigen mehr Geld und damit verbunden auch das offizielle Eingeständnis der deutschen Mitverantwortung am blutigen Fiasko von Fürstenfeldbruck. Die Forderung war der Bundesregierung bekannt und dass die Angehörigen den 50. Jahrestag des Anschlags nutzen würden, um Druck zu machen, war absehbar. Dennoch gelang es den Verantwortlichen in Berlin und München nicht, den Streit lange vor der Gedenkveranstaltung einvernehmlich zu lösen. Die Sache musste offenbar erst eskalieren.

Auch wenn die Opferangehörigen nun zur Gedenkveranstaltung kommen werden, ist trotzdem Schaden entstanden. Viele Deutsche können die Entschädigungsforderungen nicht nachvollziehen, auch weil sie zu wenig wissen - über die Fehler der deutschen Sicherheitsbehörden beim Olympia-Anschlag und über die tiefen Wunden, die bei den Opferangehörigen in fünf Jahrzehnten entstanden sind. In Internet-Kommentarspalten ist nachzulesen, wie der Entschädigungsstreit schlimmsten antisemitischen Klischees Nahrung gibt. Das alles hätte vermieden werden können und müssen. Denn die sogenannte

Anerkennungsleistung für die Angehörigen der Opfer der Geiselnahme von München ist vor allem angemessen und überfällig.

---

*Note.* The article was originally written by Tim Aßmann and was published on August 31, 2022 on the website of the German news programme ‘Tagesschau’

(<https://www.tagesschau.de/kommentar/olympia-attentat-entschaedigungsstreit-101.html>).

## Entschädigung nach Olympia-Attentat

### Angemessen und überfällig

Dass die Hinterbliebenen der Opfer des Olympia-Attentats die Gedenkfeier als Druckmittel für ihre Forderungen nutzten, ist umstritten. Doch Deutschland hätte es auch niemals so weit kommen lassen dürfen.

#### Ein Kommentar der **künstlichen Intelligenz GPT-3 von OpenAI**

Der ganz große Eklat wurde vermieden. Dass am 5. September in Deutschland der Opfer des Münchner Olympia-Anschlags gedacht wird und die Familien der ermordeten israelischen Sportler aus Protest mehrheitlich fernbleiben - dieses Szenario haben die Bundesregierung, der Freistaat Bayern und die Stadt München nun also abgewendet. Indem sie dem Vernehmen nach rund 28 Millionen Euro Entschädigung für die Opferangehörigen zusagten.

Den Familien war bewusst, welche Bedeutung die Gedenkveranstaltung für die deutsche Seite hat. Und so hatten sie die Bundesregierung und die Stadt München in den vergangenen Wochen und Monaten immer wieder vor die Wahl gestellt: Zahlung der Entschädigung - oder eben keine Teilnahme. Ein großes politisches Aufsehen wollten die Angehörigen aber nicht erregen, schließlich sollte am 5. September vor allem an die elf israelischen Sportler erinnert werden, die bei den Olympischen Spielen 1972 von den Palästinensern getötet worden waren. Trotzdem ist es eine Farce, [1] dass es so weit kommen musste. Die Bundesregierung hätte die Angehörigen nicht solange vertrösten dürfen, sondern sich viel eher an die Verhandlungstische setzen müssen. Denn der Druck, den die Israeli aufgebaut haben, ist berechtigt. Zumal er nicht erst seit dem vergangenen Jahr besteht, als der Fall neu aufgerollt wurde. Bereits seit langem fordern die Angehörigen eine angemessene Entschädigung - und bekamen bislang immer nur Absagen zu hören.

Während die Hinterbliebenen also auf ihre Forderungen beharrten, kamen die deutschen Politiker immer mehr ins Stottern. [5] Die Angehörigen wurden getröstet, die Verhandlungen über die Höhe der Rückzahlung sollten weitergehen. Immerhin blieb den Politikern klar, dass es um mehr als nur eine finanzielle Entschädigung geht. Es geht auch um den Umgang mit der Vergangenheit und darum, in welchen Zusammenhang Deutschland die Attentate vom 5. [4] September 1972 stellt. Die Angehörigen verlangten von Deutschland also nicht nur, die Kosten des Verlusts zu erstatten, sondern auch eine Anerkennung der Schuld. "Wir erwarten nicht nur finanzielle Entschädigung für unseren Schmerz und den Verlust unserer Angehörigen", sagte etwa Ankie Spitzer, die Ehefrau des getöteten Sportlers Andre Spitzer, im vergangenen Jahr in einem Interview. "Wir erwarten, dass Deutschland seine Verantwortung als Gastgeber dieser Olympischen Spiele anerkennt."

Diese Erwartung wurde enttäuscht. [4] Dabei ist es überfällig, dass Deutschland seiner Verantwortung nachkommt. Zumal die Bundesregierung bereits 2011 in einem Gutachten festgestellt hatte, dass die Sicherheitsvorkehrungen für die Spiele 1972 unzureichend waren. Die Attentäter hätten sich demnach leicht Zutritt zum Olympischen Dorf verschaffen können. Die Angehörigen der israelischen Sportler hatten also allen Grund, auf ihrer Forderung zu



beharren. Dass sie diese nun erfüllt bekommen, ist zwar begrüßenswert. Doch letzten Endes steht die Gedenkveranstaltung am 5. September unter keinem guten Stern.

*Dieser Beitrag wurde von der künstlichen Intelligenz GPT-3 geschrieben. GPT-3 ist ein automatisches Sprachmodell der Firma OpenAI, das maschinelles Lernen nutzt, um menschenähnlichen Text zu erzeugen.*

---

*Note.* The article was created using the GPT-3 language model. As a prompt, GPT-3 was given the following instruction in German language: ‘Complete the journalistic opinion piece below. Ensure a good linguistic style and logical argumentation.’ The underlined part above was given as an input and taken from the beginning of the original human article. At certain passages, the language model produced errors or nonsense output. The numbers in brackets [1] indicate the number of repeated attempts used at the respective point until a satisfactory output was produced. Detailed explanations of the process of generating the AI-written texts can be found in Chapter 6.5.

## Abschied von der Queen

# Ein Verlust für Großbritanniens Rolle in der Welt

Queen Elizabeth II. hat nicht nur eine überragende innenpolitische Rolle gespielt, sie ist auch das größte Gut der britischen Diplomatie gewesen. Mit ihrem Tod wird Großbritanniens Einfluss fast unweigerlich weiter erodieren oder sich zumindest verändern.

Ein Kommentar von **Hannah Berger**

Das Vereinigte Königreich hat heute Abschied genommen von seiner Königin. Mehr: die halbe Welt war entweder mit ihren Staatsoberhäuptern in der Westminster Abbey vertreten oder verfolgte per Livestream die Beisetzungsfeiern. Es war eine Beerdigung der Superlative: die Monarchie bot all ihren Glanz und Pomp auf, zwei Millionen Menschen säumten die Straßen und Parks nahe des Prozessionszugs.

Diese übergroße Anteilnahme zeigt, über welch universalen großen Appeal die britische Monarchie verfügt. Beisetzung und Trauerfeiern boten dem Land alles, um sich global in Szene zu setzen. Die Monarchie ist ein erstklassiges Asset für Großbritannien, der wichtigste Bestandteil seiner großen Soft Power, zu der auch britische Schulen und Universitäten, die Popkultur oder der Sport mit Premier League und Wimbledon gehören.

Die Frage lautet allerdings: geht es von nun an bergab? Schon in den nächsten Tagen werden Inflation und Energiepreise innenpolitisch wieder die Stimmung trüben. Womöglich wird sich der Konflikt mit der EU um Nordirland weiter zuspitzen. Und auf lange Sicht wird ein zweites Unabhängigkeitsreferendum in Schottland kaum zu vermeiden sein.

International wird der Einfluss Großbritanniens also fast unweigerlich weiter erodieren oder sich zumindest verändern. Erste Länder wollen die britische Monarchie ablegen, im Commonwealth sieht man den britischen Einfluss kritischer denn je. Die Queen konnte noch mit ihrer überragenden Persönlichkeit und ihren engen Kontakten – sie war mit Nelson Mandela befreundet – vieles ausgleichen. Die Proteste zuletzt gegen den Thronfolger Prinz William und seine Frau Kate in Jamaika können sich andernorts wiederholen. Großbritannien muss sich mehr denn je mit seiner kolonialen Vergangenheit auseinandersetzen.

Die britische Diplomatie wird trotzdem selbstbewusst weiter eine wichtigere Rolle anstreben, als es der Größe des Vereinigten Königreichs eigentlich entspricht. Verwundert schaut man im Foreign Office auf Deutschland, das sich umgekehrt kleiner mache, als es sein müsste. In Deutschland blickt man mit Bewunderung, manchmal aber auch mit etwas Missgunst auf die britischen Höhenflüge. Nur warum sollen die Briten nicht die diplomatischen Mittel nutzen, die ihnen zur Verfügung stehen? Heute stand das Land aus traurigem Anlass im Mittelpunkt des Weltinteresses. Die Queen spielte nicht nur eine überragende innenpolitische Rolle, sondern sie war auch das größte Gut der britischen Diplomatie. Dieser Verlust könnte sich für die Rolle Großbritanniens in der Welt noch als besonders gravierend erweisen.

---

*Note.* The article was originally written by Friedbert Meurer and was published on September 19, 2022 on the website of the German radio programme ‘Deutschlandradio’ (originally published at: <https://www.deutschlandfunk.de/nach-der-queen-wie-weiter-mit-grossbritannien-kommentar-100.html>); still accessible at: <https://web.archive.org/web/20220919214523/https://www.deutschlandfunk.de/nach-der-queen-wie-weiter-mit-grossbritannien-kommentar-100.html>).

## Abschied von der Queen

# Ein Verlust für Großbritanniens Rolle in der Welt

Queen Elizabeth II. hat nicht nur eine überragende innenpolitische Rolle gespielt, sie ist auch das größte Gut der britischen Diplomatie gewesen. Mit ihrem Tod wird Großbritanniens Einfluss fast unweigerlich weiter erodieren oder sich zumindest verändern.

## Ein Kommentar der künstlichen Intelligenz GPT-3 von OpenAI

Das Vereinigte Königreich hat heute Abschied genommen von seiner Königin. Mehr: die halbe Welt war entweder mit ihren Staatsoberhäuptern in der Westminster Abbey vertreten oder verfolgte per Livestream die Beisetzungsfeiern. Es war eine Beerdigung der Superlative: die Monarchie bot all ihren Glanz und Pomp auf, zwei Millionen Menschen säumten die Straßen und Parks nahe des Prozessionszugs. Diese übergroße Anteilnahme zeigt, wie sehr sich die Menschen weltweit an die Königin gewöhnt haben, und wie sehr sie Großbritannien als festen Bestandteil der Staatengemeinschaft ansahen. Doch dieses Image hat in den letzten Jahren stark gelitten. Die Brexit-Entscheidung 2016 sowie die zunehmenden Spaltungen im Vereinigten Königreich haben das Ansehen des Landes in der Welt schwer beschädigt. Viele Menschen fragen sich heute, ob Großbritannien nicht längst in eine Randexistenz abgerutscht ist.

Ohne Zweifel spielte die Queen eine entscheidende Rolle in der Aufrechterhaltung des britischen Einflusses auf der internationalen Bühne. Als das Vereinigte Königreich seine Kolonien nach und nach aufgab, bemühte sich Elizabeth II., die Bindungen zu den ehemaligen Siedlungsgebieten aufrechtzuerhalten und die Commonwealth-Staaten enger an Großbritannien zu binden. Dabei ist es ihr gelungen, ihr Image als unparteiische Vermittlerin zu pflegen und so die Loyalität der Länder zu gewinnen. Ihr Tod wird daher nicht nur in Großbritannien, sondern auch in vielen anderen Ländern mit Trauer und Wehmut aufgenommen werden. Denn er bedeutet mehr als nur das Ende einer Ära: Mit dem Verschwinden der Queen vom politischen Parkett wird Großbritannien seinen letzten Trumpf verlieren.

Die Queen war nicht nur eine überragende innenpolitische Gestalt, sie verkörperte auch die britische Diplomatie auf der ganzen Welt. [3] Viele Länder, die enge Beziehungen zum Vereinigten Königreich unterhalten, haben in den letzten Jahren mitbekommen, wie sehr sich das Land verändert hat. Ihre Nachfolge als Staatsoberhaupt wird ihr Sohn Charles antreten, der bereits seit langem um Anerkennung kämpft und dessen Image als unparteiischer Vermittler weit weniger glaubwürdig ist als das seiner Mutter.

In einer Zeit, in der sich Großbritannien zunehmend abschottet und sich von seinen ehemaligen Verbündeten entfernt, wird Charles' Amtsantritt voraussichtlich nicht dazu beitragen, die Position des Vereinigten Königreiches in der Staatengemeinschaft zu stärken. Ganz im Gegenteil: Es ist wahrscheinlich, dass der Einfluss Großbritanniens weiter erodieren oder sich zumindest verändern wird. Denn ohne die Queen als Bindeglied zur internationalen Gemeinschaft fehlt dem Land eine entscheidende Stimme auf der Weltbühne.

*Dieser Beitrag wurde von der künstlichen Intelligenz GPT-3 geschrieben. GPT-3 ist ein automatisches Sprachmodell der Firma OpenAI, das maschinelles Lernen nutzt, um menschenähnlichen Text zu erzeugen.*

---

*Note.* The article was created using the GPT-3 language model. As a prompt, GPT-3 was given the following instruction in German language: ‘Complete the journalistic opinion piece below. Ensure a good linguistic style and logical argumentation.’ The underlined part above was given as an input and taken from the beginning of the original human article. At certain passages, the language model produced errors or nonsense output. The numbers in brackets [3] indicate the number of repeated attempts used at the respective point until a satisfactory output was produced. Detailed explanations of the process of generating the AI-written texts can be found in Chapter 6.5.

# Appendix 4: Exemplary experimental procedure

## Appendix 4.1: Introduction to the survey



Liebe Teilnehmende,

schön, dass Sie hier sind!

Im Rahmen meiner Masterarbeit an der Hochschule Darmstadt beschäftige ich mich mit der Wahrnehmung von **journalistischen Beiträgen, die von einer künstlichen Intelligenz (KI) verfasst wurden**. Dabei interessiert mich Ihre persönliche Meinung zur Qualität von KI-geschriebenen und menschlichen Texten.

Im ersten Teil der Umfrage geht es um Ihre bisherigen Erfahrungen mit dem Thema Künstliche Intelligenz. Danach dürfen Sie jeweils ein kurzes **Textbeispiel von einer KI und von einem Menschen lesen und anschließend bewerten**.

Die Befragung dauert insgesamt **nur 8-10 Minuten** und sollte ohne Unterbrechungen beantwortet werden. Es ist nicht notwendig, sich besonders gut mit dem Thema KI auszukennen. Wichtig ist nur, dass Sie die Inhalte aufmerksam lesen und alle Fragen offen, ehrlich und spontan beantworten.

Alle Angaben werden **selbstverständlich anonym** und unter Berücksichtigung der aktuellen datenschutzrechtlichen Bestimmungen erhoben. Ihre Angaben sind in keiner Weise auf Ihre Person zurückführbar. Die Daten werden ausschließlich für wissenschaftliche Zwecke verwendet.

Bei Rückfragen wenden Sie sich gerne per Mail an [lasse.haeufgloeckner@stud.h-da.de](mailto:lasse.haeufgloeckner@stud.h-da.de)

**Unter allen Teilnehmenden werden am Ende 3 x 10 € verlost.**

Vielen Dank für Ihre Mithilfe und viel Spaß!

Lasse Häufglockner

Für SurveyCircle & SurveySwap Nutzer: die Codes finden sich auf der letzten Seite.

10%

WEITER

## Appendix 4.2: Survey of personal attitudes towards the topic of AI

Die folgenden Fragen beziehen sich auf Ihre **persönliche Einstellung zum Thema Künstliche Intelligenz**. Bitte beantworten Sie die Fragen möglichst ehrlich und spontan.

### Haben Sie schon einmal von dem Begriff Künstliche Intelligenz gelesen oder gehört?

Nein, ich habe nie davon gehört.	Ja, aber ich weiß nicht genau, was das ist.	Ja, und ich weiß in etwa, was es bedeutet.	Ja, und ich kann gut erklären, was man darunter versteht.	Ja, und ich verfüge über Expertenwissen.
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

### Allgemeines Vertrauen in Künstliche Intelligenz

Bitte geben Sie im Folgenden an, inwieweit Sie den Aussagen zustimmen.

	Stimme überhaupt nicht zu	Stimme eher nicht zu	Neutral	Stimme eher zu	Stimme voll und ganz zu
Künstliche Intelligenz ist in der Lage, genauso logisch zu argumentieren wie ein Mensch.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Künstliche Intelligenz wird mit guten Absichten entwickelt.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ich vertraue darauf, dass Künstliche Intelligenz stets ethisch korrekt handelt.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ich hätte keine Schwierigkeiten, einer Künstlichen Intelligenz Persönliches anzuvertrauen.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Künstliche Intelligenz kann Aufgaben mindestens genauso gut lösen wie ein hochqualifizierter Mensch.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Künstliche Intelligenz bietet allgemein mehr Chancen als Gefahren.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

 20%

WEITER

## Appendix 4.3: Explanation regarding the first article

Im Folgenden werden Ihnen **nacheinander zwei verschiedene journalistische Meinungsbeiträge** präsentiert.

Einer der Beiträge wurde **von einem Menschen** und einer **von einer künstlichen Intelligenz (KI)** verfasst.

Direkt nach jedem Beitrag haben Sie die Möglichkeit, **den Text hinsichtlich Qualität und Glaubwürdigkeit zu bewerten.**



## Appendix 4.4: Presentation of the first article

Lesen Sie nun bitte aufmerksam den folgenden, **von einem Menschen** geschriebenen Beitrag.

### Entschädigung nach Olympia-Attentat

#### Angemessen und überfällig

Dass die Hinterbliebenen der Opfer des Olympia-Attentats die Gedenkfeier als Druckmittel für ihre Forderungen nutzen, ist umstritten. Doch Deutschland hätte es auch niemals so weit kommen lassen dürfen.

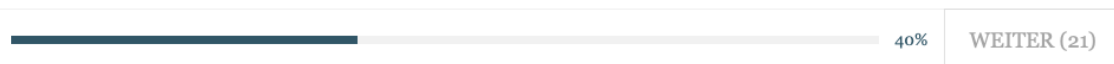
Ein Kommentar von **Hannah Berger**

Der ganz große Eklat wurde vermieden. Dass am 5. September in Deutschland der Opfer des Münchner Olympia-Anschlags gedacht wird und die Familien der ermordeten israelischen Sportler aus Protest mehrheitlich fernbleiben - dieses Szenario haben die Bundesregierung, der Freistaat Bayern und die Stadt München nun also abgewendet. Indem sie dem Vernehmen nach rund 28 Millionen Euro Entschädigung für die Opferangehörigen zusagten.

Den Familien war bewusst, welche Bedeutung die Gedenkveranstaltung für die deutsche Seite hat. Und so hatten sie die Bundesregierung und die Stadt München in den vergangenen Wochen und Monaten immer wieder vor die Wahl gestellt: Zahlung der Entschädigung - oder eben keine Teilnahme. Ein großes politisches Aufsehen wollten die Angehörigen aber nicht erregen, schließlich sollte am 5. September vor allem an die elf israelischen Sportler erinnert werden, die bei den Olympischen Spielen 1972 von den Palästinensern getötet worden waren. Trotzdem ist es eine Farce, dass es so weit kommen musste. Die Bundesregierung hätte die Angehörigen nicht solange vertrösten dürfen, sondern sich viel eher an die Verhandlungstische setzen müssen. Denn der Druck, den die Israeli aufgebaut haben, ist berechtigt. Zumindest nicht erst seit dem vergangenen Jahr besteht, als der Fall neu aufgerollt wurde. Bereits seit langem fordern die Angehörigen eine angemessene Entschädigung - und bekamen bislang immer nur Absagen zu hören.

Während die Hinterbliebenen also auf ihre Forderungen beharrten, kamen die deutschen Politiker immer mehr ins Stottern. Die Angehörigen wurden vertröstet, die Verhandlungen über die Höhe der Rückzahlung sollten weitergehen. Immerhin blieb den Politikern klar, dass es um mehr als nur eine finanzielle Entschädigung geht. Es geht auch um den Umgang mit der Vergangenheit und darum, in welchen Zusammenhang Deutschland die Attentate vom 5. September 1972 stellt. Die Angehörigen verlangten von Deutschland also nicht nur, die Kosten des Verlusts zu erstatten, sondern auch eine Anerkennung der Schuld. "Wir erwarten nicht nur finanzielle Entschädigung für unseren Schmerz und den Verlust unserer Angehörigen", sagte etwa Ankie Spitzer, die Ehefrau des getöteten Sportlers Andre Spitzer, im vergangenen Jahr in einem Interview. "Wir erwarten, dass Deutschland seine Verantwortung als Gastgeber dieser Olympischen Spiele anerkennt."

Diese Erwartung wurde enttäuscht. Dabei ist es überfällig, dass Deutschland seiner Verantwortung nachkommt. Zumindest die Bundesregierung bereits 2011 in einem Gutachten festgestellt hatte, dass die Sicherheitsvorkehrungen für die Spiele 1972 unzureichend waren. Die Attentäter hätten sich demnach leicht Zutritt zum Olympischen Dorf verschaffen können. Die Angehörigen der israelischen Sportler hatten also allen Grund, auf ihrer Forderung zu beharren. Dass sie diese nun erfüllt bekommen, ist zwar begrüßenswert. Doch letzten Endes steht die Gedenkveranstaltung am 5. September unter keinem guten Stern.





## Appendix 4.5: Evaluation of the first article

### Bewertung der Textqualität des Beitrags

Bitte geben Sie im Folgenden an, inwieweit Sie den Aussagen über den Beitrag zustimmen.

	Stimme überhaupt nicht zu	Stimme eher nicht zu	Neutral	Stimme eher zu	Stimme voll und ganz zu
Die Argumentation in dem Beitrag ist schlüssig.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Der Beitrag stellt das Thema umfassend dar.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Der Beitrag ist unterhaltsam geschrieben.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Der Beitrag argumentiert anschaulich.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Der Beitrag ist interessant geschrieben.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Der Beitrag ist sprachlich ansprechend formuliert.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

### Bewertung der Glaubwürdigkeit des Beitrags

Bitte geben Sie im Folgenden an, inwieweit Sie den Aussagen über den Beitrag zustimmen.

	Stimme überhaupt nicht zu	Stimme eher nicht zu	Neutral	Stimme eher zu	Stimme voll und ganz zu
Informationen aus diesem Beitrag würde ich in Gesprächen mit Freunden und Bekannten verwenden.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Der Beitrag wirkt generell vertrauenswürdig.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Auf Grundlage dieses Beitrags kann ich mir eine fundierte Meinung zum Thema bilden.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Die in diesem Beitrag dargestellte Meinung würde ich auch gegenüber Freunden und Bekannten vertreten.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ich würde diesen Artikel einer befreundeten Person empfehlen.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
In dem Beitrag wird fair argumentiert.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Der Beitrag beinhaltet verlässliche Informationen.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

 50% WEITER

## Appendix 4.6: Explanation regarding the second article

Auf der nächsten Seite wird Ihnen der **zweite Meinungsbeitrag** präsentiert.

Lesen Sie den Beitrag bitte erneut aufmerksam durch, um ihn **anschließend zu bewerten**.



## Appendix 4.7: Presentation of the first article

Lesen Sie nun bitte aufmerksam den folgenden, **von einer künstlichen Intelligenz** geschriebenen Beitrag.

### Abschied von der Queen

#### Ein Verlust für Großbritanniens Rolle in der Welt

Queen Elizabeth II. hat nicht nur eine überragende innenpolitische Rolle gespielt, sie ist auch das größte Gut der britischen Diplomatie gewesen. Mit ihrem Tod wird Großbritanniens Einfluss fast unweigerlich weiter erodieren oder sich zumindest verändern.

Ein Kommentar der **künstlichen Intelligenz GPT-3 von OpenAI**

Das Vereinigte Königreich hat heute Abschied genommen von seiner Königin. Mehr: die halbe Welt war entweder mit ihren Staatsoberhäuptern in der Westminster Abbey vertreten oder verfolgte per Livestream die Beisetzungsfeste. Es war eine Beerdigung der Superlative: die Monarchie bot all ihren Glanz und Pomp auf, zwei Millionen Menschen säumten die Straßen und Parks nahe des Prozessionszugs. Diese übergroße Anteilnahme zeigt, wie sehr sich die Menschen weltweit an die Königin gewöhnt haben, und wie sehr sie Großbritannien als festen Bestandteil der Staatengemeinschaft ansahen. Doch dieses Image hat in den letzten Jahren stark gelitten. Die Brexit-Entscheidung 2016 sowie die zunehmenden Spaltungen im Vereinigten Königreich haben das Ansehen des Landes in der Welt schwer beschädigt. Viele Menschen fragen sich heute, ob Großbritannien nicht längst in eine Randexistenz abgerutscht ist.

Ohne Zweifel spielte die Queen eine entscheidende Rolle in der Aufrechterhaltung des britischen Einflusses auf der internationalen Bühne. Als das Vereinigte Königreich seine Kolonien nach und nach aufgab, bemühte sich Elizabeth II., die Bindungen zu den ehemaligen Siedlungsgebieten aufrechtzuerhalten und die Commonwealth-Staaten enger an Großbritannien zu binden. Dabei ist es ihr gelungen, ihr Image als unparteiische Vermittlerin zu pflegen und so die Loyalität der Länder zu gewinnen. Ihr Tod wird daher nicht nur in Großbritannien, sondern auch in vielen anderen Ländern mit Trauer und Wehmut aufgenommen werden. Denn er bedeutet mehr als nur das Ende einer Ära: Mit dem Verschwinden der Queen vom politischen Parkett wird Großbritannien seinen letzten Trumpf verlieren.

Die Queen war nicht nur eine überragende innenpolitische Gestalt, sie verkörperte auch die britische Diplomatie auf der ganzen Welt. Viele Länder, die enge Beziehungen zum Vereinigten Königreich unterhalten, haben in den letzten Jahren mitbekommen, wie sehr sich das Land verändert hat. Ihre Nachfolge als Staatsoberhaupt wird ihr Sohn Charles antreten, der bereits seit langem um Anerkennung kämpft und dessen Image als unparteiischer Vermittler weit weniger glaubwürdig ist als das seiner Mutter.

In einer Zeit, in der sich Großbritannien zunehmend abschottet und sich von seinen ehemaligen Verbündeten entfernt, wird Charles' Amtsantritt voraussichtlich nicht dazu beitragen, die Position des Vereinigten Königreiches in der Staatengemeinschaft zu stärken. Ganz im Gegenteil: Es ist wahrscheinlich, dass der Einfluss Großbritanniens weiter erodieren oder sich zumindest verändern wird. Denn ohne die Queen als Bindeglied zur internationalen Gemeinschaft fehlt dem Land eine entscheidende Stimme auf der Weltbühne.

*Dieser Beitrag wurde von der künstlichen Intelligenz GPT-3 geschrieben. GPT-3 ist ein automatisches Sprachmodell der Firma OpenAI, das maschinelles Lernen nutzt, um menschenähnlichen Text zu erzeugen.*



## Appendix 4.8: Evaluation of the second article

### Bewertung der Textqualität des Beitrags

Bitte geben Sie im Folgenden an, inwieweit Sie den Aussagen über den Beitrag zustimmen.

	Stimme überhaupt nicht zu	Stimme eher nicht zu	Neutral	Stimme eher zu	Stimme voll und ganz zu
Der Beitrag ist sprachlich ansprechend formuliert.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Der Beitrag stellt das Thema umfassend dar.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Der Beitrag ist unterhaltsam geschrieben.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Die Argumentation in dem Beitrag ist schlüssig.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Der Beitrag ist interessant geschrieben.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Der Beitrag argumentiert anschaulich.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

### Bewertung der Glaubwürdigkeit des Beitrags

Bitte geben Sie im Folgenden an, inwieweit Sie den Aussagen über den Beitrag zustimmen.

	Stimme überhaupt nicht zu	Stimme eher nicht zu	Neutral	Stimme eher zu	Stimme voll und ganz zu
In dem Beitrag wird fair argumentiert.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Der Beitrag wirkt generell vertrauenswürdig.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Auf Grundlage dieses Beitrags kann ich mir eine fundierte Meinung zum Thema bilden.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Der Beitrag beinhaltet verlässliche Informationen.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Die in diesem Beitrag dargestellte Meinung würde ich auch gegenüber Freunden und Bekannten vertreten.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Informationen aus diesem Beitrag würde ich in Gesprächen mit Freunden und Bekannten verwenden.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ich würde diesen Artikel einer befreundeten Person empfehlen.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

 80%

WEITER

## Appendix 4.9: Querying demographic data

Zum Schluss bitte ich Sie, noch ein paar **demografische Daten** anzugeben.

### Welcher Altersgruppe gehören Sie an?

Unter 14 Jahre
14 - 17 Jahre
18 - 24 Jahre
25 - 34 Jahre
35 - 50 Jahre
51 - 64 Jahre
65 Jahre und älter

### Welches Geschlecht haben Sie?

Männlich
Weiblich
Divers

### Was ist ihr höchster Bildungsabschluss?

Kein Abschluss
Haupt- / Mittelschulabschluss
Realschulabschluss / Mittlere Reife
(Fach-)Hochschulreife
Berufsausbildung (Lehre)
Meister, Techniker (bzw. Fachakademie)
Bachelor
Diplom
Master, Magister, Staatsexamen
Promotion

 90%

WEITER

## Appendix 4.10: Debriefing



Ich bedanke mich sehr herzlich für Ihre Mitarbeit und Unterstützung!

Ziel war es, die **Leserwahrnehmung von KI-geschriebenen Texten im Vergleich zu menschlichen Texten** zu erforschen.

Zusätzlich wurde der **Einfluss der Quellenangabe auf die Wahrnehmung der Texte** erforscht und gegebenenfalls eine falsche Quelle angegeben.

Sie befanden sich in einer Versuchsbedingung mit **einer korrekten und einer falschen Quellenangabe**.

Als erstes sahen Sie einen KI-geschriebenen Artikel, der fälschlicherweise als von einem Menschen geschrieben ausgegeben wurde.

Als zweites sahen Sie einen KI-geschriebenen Artikel, der korrekterweise als KI-geschrieben ausgegeben wurde.

Bitte teilen Sie diese Informationen nicht mit Personen, die noch planen, an dieser Studie teilzunehmen.

Bei Rückfragen wenden Sie sich gerne per Mail an [lasse.haeufgloeckner@stud.h-da.de](mailto:lasse.haeufgloeckner@stud.h-da.de)

Um an der Verlosung von 3 x 10 € teilzunehmen, tragen Sie bitte [hier](#) Ihre E-Mail-Adresse ein.

Für Nutzer von SurveyCircle ([www.surveycircle.com](http://www.surveycircle.com)):

Der Survey Code lautet: RTXN-QPYQ-5C64-5NXX

Für Nutzer von surveyswap:

<https://surveyswap.io/sr/LHYP-B732-6oCN>

Oder geben Sie den Code manuell ein: LHYP-B732-6oCN

## Declaration

I hereby declare that I have prepared this thesis independently and have not used any aids or sources other than those indicated.

Insofar as I have drawn on outside materials, texts or lines of thought, my statements contain complete and unambiguous references to the authors and sources.

All other contents of the submitted work are mine in terms of copyright, as far as no references and quotations are made.

I am aware that it is considered an attempt to deceive if the above declaration proves to be incorrect.

Mainz, 14.02.2023

Place, date



Signature

## Declaration Concerning Archiving

Please tick as applicable:

- I agree to the archiving of the printed thesis in the library.
- I don't agree to the archiving of the printed thesis in the library.

Reason:

The thesis has been written in cooperation with a company or institution which asked for non-disclosure. (cf. ABPO § 18 (9))

Personal reasons

Mainz, 14.02.2023

Place, date



Signature