Perceptions of AI in Science Communication

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Angelica Henestrosa
geb. Lermann Henestrosa
aus Mexiko-Stadt/Mexiko

Tübingen
2025

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	09.07.2025
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Apl. Prof. Dr. Joachim Kimmerle
2. Berichterstatterin:	Prof. Dr. Sonja Utz

Para mamá

Contents

Summary

In today's information society, anyone with Internet access theoretically has a gateway to a wealth of information. There, individuals seeking knowledge can access a diverse array of information from numerous sources. However, this abundance increases the need for reliable and verified content. This is especially true for communicating scientific knowledge, which is indispensable for overcoming the multiple crises of our time. Generative artificial intelligence (GenAI) has recently been added to possible sources of information, holding both opportunities and risks for science communication. On the one hand, it has the potential to revolutionize the dialogue between science and society; on the other hand, the underlying technology poses a challenge to the established principles of scientific validation and verification. What is certain is that artificial intelligence (AI) is and will be used to communicate science, retrieve scientific information, and generate scientific knowledge. As earlier approaches to communicating content via AI fundamentally differ from today's Large Language Models (LLMs), little is known about how humans perceive AI authorship of more complex topics and texts.

This dissertation contribute to filling this research gap by using experimental and survey studies to investigate how people perceive AI authorship, what attitudes and concepts they hold about it, and whether information about AI influences these perceptions. The results show little difference in the credibility and trustworthiness between AI and human authors. At the same time, attitudes toward AI tend to be positive, but knowledge is scarce, and concepts about this new technology are vague. Moreover, providing information on the strengths or limitations of AI, for example, hardly influences the evaluation of AI-generated texts and the perception of AI authorship. In sum, these findings indicate that readers of science communication information may exhibit indifference toward the authorship of the content. By revealing this agnosticism, this work contributes to the growing body of research on the perception of AI-generated output and the factors influencing this perception. The results emphasize the need for interventions to promote the safe use of AI in science communication and the responsibility of policy makers and providers to design frameworks that take into account the capabilities and risks of GenAI.

Zusammenfassung

In der heutigen Informationsgesellschaft hat theoretisch jede und jeder mit Internetanschluss Zugang zu einer Fülle von Informationen. Wissbegierige können dort auf eine Vielzahl von Informationen aus zahlreichen Quellen zugreifen. Mit der Fülle steigt jedoch auch der Bedarf an vertrauenswürdigen und geprüften Inhalten. Dies gilt insbesondere für die Vermittlung von wissenschaftlichen Erkenntnissen, die für die Bewältigung der vielfältigen Krisen unserer Zeit unerlässlich sind. Zu den möglichen Informationsquellen ist nun Generative Künstliche Intelligenz (GenAI) hinzugekommen, die sowohl Chancen als auch Risiken für die Wissenschaftskommunikation birgt. Einerseits verfügt sie über das Potenzial, den Dialog zwischen Wissenschaft und Gesellschaft zu revolutionieren, andererseits stellt die zugrunde liegende Technologie eine Herausforderung für die etablierten Grundsätze wissenschaftlicher Validierung und Verifizierung dar. Sicher ist, dass Künstliche Intelligenz (KI) eingesetzt wird, um Wissenschaft zu vermitteln, wissenschaftliche Informationen abzurufen und wissenschaftliche Erkenntnisse zu gewinnen. Da sich frühere Ansätze zur Kommunikation von Inhalten durch KI grundlegend von heutigen Large Language Models (LLMs) unterscheiden, ist wenig darüber bekannt, wie Menschen KI-Autorenschaft bei komplexeren Themen und Texten wahrnehmen.

Diese Dissertation trägt zur Schließung dieser Forschungslücke bei, indem sie anhand von experimentellen und Umfragestudien untersucht, wie Menschen KI-Autorenschaft wahrnehmen, welche Einstellungen und Konzepte sie dazu haben und ob Informationen über KI diese Wahrnehmungen beeinflussen. Die Ergebnisse zeigen kaum Unterschiede in der Glaubwürdigkeit und Vertrauenswürdigkeit zwischen KI und menschlichen Autoren. Gleichzeitig sind die Einstellungen gegenüber KI tendenziell positiv, aber das Wissen über diese neue Technologie ist gering, und die Vorstellungen sind vage. Darüber hinaus beeinflussen Vorabinformationen, beispielsweise über die Stärken oder Schwächen der KI, die Bewertung von KI-generierten Texten und die Wahrnehmung der KI-Autorenschaft kaum. Insgesamt deuten diese Ergebnisse darauf hin, dass Leserinnen und Leser von Informationen aus dem Bereich der Wissenschaftskommunikation der Autorenschaft des Inhalts gegenüber gleichgültig sein könnten. Durch die Aufdeckung dieses Agnostizismus trägt diese Arbeit zur wachsenden Zahl von Forschungsarbeiten über die Wahrnehmung KI-generierter Inhalte und den Faktoren, die diese Wahrnehmung beeinflussen, bei. Die Ergebnisse unterstreichen den Bedarf an Maßnahmen zur Förderung des sicheren Einsatzes von KI in der Wissenschaftskommunikation und die Verantwortung von politischen Entscheidungsträgerinnen und Anbietern, Rahmenbedingungen zu schaffen, die den Fähigkeiten und Risiken von GenAI gerecht werden.

Lay Summary

This lay summary was generated with Copilot version 4.0 on April 9, 2025, using the prompt "Prepare a short lay summary based on this summary of a dissertation: [Summary]". I have carefully reviewed the content and take full responsibility for its accuracy:

In today's digital age, the internet offers vast amounts of information, making it crucial to ensure the reliability of content, especially in science communication. Generative artificial intelligence (GenAI) has emerged as a new source of information, presenting both opportunities and challenges. While GenAI can enhance the interaction between science and society, it also raises concerns about maintaining scientific standards. This dissertation explores how people perceive AI authorship of complex topics. Through experiments and surveys, it finds that people generally trust AI-generated content as much as human-written content. Although attitudes towards AI are positive, there is a lack of understanding about the technology. Providing information about AI's strengths and limitations does not significantly change people's perceptions. Overall, the study reveals that readers are indifferent to whether content is authored by AI or humans. This highlights the need for policies and frameworks to ensure the safe use of AI in science communication, considering both its potential and risks.

Contributions

This cumulative dissertation is based on four manuscripts and one data descriptor, all of which have been published in peer-reviewed journals. Appendix B contains copies of the five papers. In the following, the individual contributions of the authors involved in each manuscript are stated and accompanied by information on the corresponding Study Package.

Manuscript I

Lermann Henestrosa, A., Greving, H., & Kimmerle, J. (2023). Automated Journalism: The Effects of AI Authorship and Evaluative Information on the Perception of a Science Journalism Article. *Computers in Human Behavior, 138*, Article 107445. https://doi.org/10.1016/j.chb.2022.107445

Author contributions:

Author	Scientific Ideas	Data Generation	Analysis & Interpretation	Paper Writing
Lermann Henestrosa, A.	80%	100%	80%	70%
Greving, H.	0%	0%	10%	15%
Kimmerle, J.	20%	0%	10%	15%

Status in publication process: Published

Part of Study Package 1: Evaluative AI: Credibility and Trust Perceptions in Science Journalism

Manuscript II

Lermann Henestrosa, A., & Kimmerle, J. (2024). The Effects of Assumed AI vs. Human Authorship on the Perception of a GPT-generated Text. *Journalism and Media, 5*(3), 1085-1097. https://doi.org/10.3390/journalmedia5030069

Author contributions:

Author	Scientific Ideas	Data Generation	Analysis & Interpretation	Paper Writing
Lermann Henestrosa, A.	80%	100%	90%	80%
Kimmerle, J.	20%	0%	10%	20%

Status in publication process: Published

Part of Study Package 1: Evaluative AI: Credibility and Trust Perceptions in Science Journalism

Manuscript III

Lermann Henestrosa, A., & Kimmerle, J. (2024). Understanding and Perception of Automated Text Generation among the Public: Two Surveys with Representative Samples in Germany. *Behavioral Sciences, 14*(5), Article 353. https://doi.org/10.3390/bs14050353

Author contributions:

Author	Scientific Ideas	Data Generation	Analysis & Interpretation	Paper Writing
Lermann Henestrosa, A.	90%	100%	90%	80%
Kimmerle, J.	10%	0%	10%	20%

Status in publication process: Published

Part of Study Package 2: Attitudes toward and Perceptions of Automated Text Generation

Manuscript IV

Lermann Henestrosa, A., & Kimmerle, J. (2024). Data Descriptor for "Understanding and Perception of Automated Text Generation among the Public: Two Surveys with Representative Samples in Germany". *Data, 9*(10), Article 116. https://doi.org/10.3390/data9100116

Author contributions:

Author	Scientific Ideas	Data Generation	Analysis & Interpretation	Paper Writing
Lermann Henestrosa, A.	90%	-	-	90%
Kimmerle, J.	10%	-	-	10%

Status in publication process: Published

Part of Study Package 2: Attitudes toward and Perceptions of Automated Text Generation

Manuscript V

Lermann Henestrosa, A., & Kimmerle, J. (2025). "Always Check Important Information!"—The Role of Disclaimers in the Perception of AI-generated Content. *Computers in Human Behavior: Artificial Humans, 4*, Article 100142. https://doi.org/10.1016/j.chbah.2025.100142

Author contributions:

Author	Scientific Ideas	Data Generation	Analysis & Interpretation	Paper Writing
Lermann Henestrosa, A.	80%	100%	80%	80%
Kimmerle, J.	20%	0%	20%	20%

Status in publication process: Published

Part of Study Package 3: The Effects of Disclaimers on Readers' Perceptions of AI Output

1 Introduction

"We've learned to make machines that can mindlessly generate text, but we haven't learned how to stop imagining the mind behind it."—Emily M. Bender[1]

Over thousands of years, various written languages have developed into the unique medium of communication and repository of knowledge that we know today. They are a cultural achievement that requires cognitive, social-organizational skills as well as technological developments. While the ability to speak has a biological basis and is closely linked to human evolution (De Boer, 2017), the ability to write is anthropogenic. However, the belief that the ability to write and communicate through writing is reserved for humans may have been shattered for many in recent years. This in no way refers to the ability to write with a pen on paper, as robotic arms have been able to do this for some time. Nor is it the ability to hold simple conversations, as we have long known from chatbots, or to automatically fill in templates, as has been done in automated journalism for over 10 years. What was demonstrated in November 2022 with the publication of ChatGPT—the first generally accessible Large Language Model (LLM)—is the ability of machines to produce fluent and, above all, meaningful-sounding written texts without any human intervention, apart from simple user instructions.

Although it is not the first skill initially reserved for humans that were taught machines (e.g., driving a car or playing chess), this type of Artificial Intelligence (AI) holds particular potential, as text and writing are central to many areas of life. Automatically generated output on almost any conceivable input, which is no longer distinguishable from human creation (Köbis & Mossink, 2021), is about to fundamentally change how we communicate, learn, work, and access knowledge. Therefore, enthusiasm and fear have characterized the debates surrounding this technology since the publication of LLMs. However, the first attempts to have computers write were already made in the middle of the last century.

Back in the 1960s, Weizenbaum succeeded in simulating a written conversation, which amazed the people who were interacting with the computer called ELIZA (Weizenbaum, 1966). However, these early approaches to Automated Text Generation (ATG) were those of maximum control, in which the programmer implemented rules to determine the machine's output—context understanding and flexible adaptation were not yet possible. Today's Language Models (LMs) are based on a different approach. First, the additional "large" stands for an enormous amount of data, which is used to train the models (Shanahan, 2024). On the one hand, this training makes linguistically outstanding performance

[1] https://nymag.com/intelligencer/article/ai-artificial-intelligence-chatbots-emily-m-bender.html [accessed April 1, 2025]

possible in the first place, but impeding at the same time reliable curation and fact-based control of content because of the large amount of data that is needed for it and the uncontrollability of Neural Networks (NNs; Shen et al., 2023; Zhao et al., 2025). Moreover, the underlying training on string prediction tasks of LMs, which aims at predicting the probability of a token, that is, a string or word (Bender et al., 2021; Ooi et al., 2025), illustrates both the potential and pitfalls of products based on this technology: The produced text might be flawless and persuasive on a purely linguistic level, while not necessarily being factually accurate (McGowan et al., 2023). This presents a significant challenge for individuals seeking reliable scientific information, especially those who intend to use LLMs or encounter AI-generated content on such topics, whether voluntarily or involuntarily.

In a nutshell, ChatGPT paved the way for a technology the public has broad access to, which, as is often the case, has enormous potential but also specific risks (for an overview, see Weidinger et al., 2022). However, the sudden widespread availability confronts the public with challenges whose handling lags the technological development. Research is needed on how AI authorship is perceived, what people think about it, and how a responsible approach to GenAI can be shaped. This serves both the informed evaluation of the opportunities and risks of AI, which can produce text on any possible topic, and the safe and effective use of this technology.

1.1 Introducing AI

When talking about AI, I refer to an umbrella term introduced in the 1950s during a workshop on the simulation of human learning by machines (McCarthy et al., 2006). The assumption at the time was that "every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it" (as cited in McCarthy et al., 2006, p. 13). As bold as these and subsequent lines may sound in retrospect, the basic idea is still relevant and drives today's AI research, which is only possible because understanding is not necessary to remodel human cognitive performances.

Phases of stagnation and breakthroughs have characterized the development of AI ever since. In addition to successes with so-called "expert systems" that outperformed humans in specific tasks, such as playing board games (e.g., AlphaGo), other essential milestones in AI development include the introduction of NNs in the 1980s, the development of Deep Learning (DL) in the 2010s (Goodfellow et al., 2016; Hinton et al., 2012), and the Transformer Architecture in 2017 (Vaswani et al., 2017), which forms the basis for modern LLMs. Since an outline of the development history of AI up to today's LMs is beyond the scope of this dissertation, please refer to overview works (e.g., Haenlein & Kaplan, 2019).

In the meantime, AI has become an inflationary buzzword for technology and digi-

tization, making it difficult to clearly define and distinguish it from technologies based on algorithms. Whereas an algorithm is often described as a step-by-step formula like a recipe (e.g., Misselhorn, 2019, p. 19), AI goes beyond this definition. On the one hand, different techniques are used to enable learning and adaptation; on the other hand, AI systems can process vast amounts of data to make decisions and predictions that are not explicitly programmed. AI can refer to a subfield of computer science or also, more specifically, to "machines that act rationally" (Poole et al., 1998; Russell & Norvig, 2016), as well as various methods such as Machine Learning (ML) or Natural Language Processing (NLP), which are often intertwined. A recurring distinction is that of weak and strong AI, which is based on the Chinese Room Argument (Searle, 1980, 1990) and describes weak AI as a simulation of human thought, while strong AI is described as actual thought (Russell & Norvig, 2016). Given the extensive literature on finding a definition of AI (e.g., P. Wang, 2019) and the resulting technical and philosophical implications (McCarthy & Hayes, 1981; Rosengrün, 2021, e.g.,), I will not propose a new definition. Instead, in this dissertation, I generally refer to GenAI and LLMs as to how the most famous tool describes itself[2]:

"Generative AI refers to a class of artificial intelligence systems designed to create new content—such as text, images, audio, video, or code—based on patterns learned from existing data. It uses models trained on large datasets to generate outputs that resemble human-created content. Large Language Models are a type of generative AI specifically trained to understand and produce human language. They are deep learning models, often based on transformer architectures, trained on vast amounts of text data to perform tasks like text generation, translation, summarization, and question answering."

These resulting products have in common that they were traditionally the achievement of human creation and effort and can now be produced automatically. However, the development of LLMs began long before ChatGPT, Llama, Claude, and other LLMs were released. Different approaches were used to automate writing processes and make the production of boilerplate content more efficient. These included templates to automatically fill in recurring texts (e.g., reports in journalism, product descriptions) or the training of smaller models with small data sets. The limitations of these approaches lay in the lack of availability of structured data (Graefe et al., 2016), which would have made it possible to extend them to other subject areas. As a result, the use of ATG tended to be limited to supporting specific, writing-heavy areas such as data journalism and short news reporting, meaning that its potential for other application areas was underestimated (Schäfer, 2023; Tatalovic, 2018).

Recent advancements in NLP and the availability of powerful computing resources have enabled the training of complex models. The launch of the Transformer Architecture

[2]This definition was generated with ChatGPT 4o on April 4, 2025. The following prompt was used: "What is GenAI and what are LLMs? Provide a concise and complete definition of both terms."

(Vaswani et al., 2017) specifically meant the introduction of the attention mechanism, which allows relevant parts of a text to be weighted instead of going sequentially from one word to the next to capture the meaning of an entire text (Brown et al., 2020). Above all, this led to efficient processing of large text data and better context recognition. Although it had been possible to consult voice assistants such as Siri or Alexa for some time, their answers lacked quality, precision, and breadth, not least due to limited processing capacities and a lack of contextual understanding (Mahmood et al., 2025). With the subsequent training of LMs on an enormous number of parameters (Abukmeil et al., 2022), the way is paved for LLMs and makes it possible to generate text on almost any conceivable topic.

Currently, the implications for science communication remain relatively ambiguous when considering the ability of LLMs to generate texts also on scientific topics within seconds. For instance, everyone could sit down and finally have the string theory explained to them in simple terms with the additional option of asking 100 times. However, there are high standards for the accuracy of scientific information, which can be improved through further training runs (OpenAI et al., 2023). Absolute adherence to facts, however, is difficult to address simply due to the mass of training data in the model itself, though real-time search and fact checks, for example, may contribute.

Traditionally, other areas in which AI systems play a significant role, such as automated driving or facial recognition, are the focus of public debate for quite some time before their broad market launch. This allows for opinions on the opportunities and risks to be formed beforehand and to address potential security problems at an early stage. In the case of LLMs, attention was drawn simultaneously as they were made widely available. Consequently, since the use of ATG was limited to very few areas before the release of ChatGPT and exploded virtually overnight in all areas, including scientific content and science communication, there is now a fundamental lack of research on the perception and interpretation of such content and AI authorship.

1.2 Science Communication in the Era of GenAI

Science communication, involving a diverse array of stakeholder, is a domain that will undoubtedly be impacted by GenAI. Both clear benefits and emerging challenges due to the problems mentioned above are conceivable.

In a post-factual era with global political movements increasingly questioning the authority of science (Bijker et al., 2009), the public relies more than ever on safe and effective science communication. Despite the theoretically unrestricted access to scientific information via the internet, obtaining sound and reliable scientific facts is not as easy as it seems. Although there is an increasing trend toward opening research practices

(Committee on Toward an Open Science Enterprise et al., 2018; Laakso et al., 2011), the scientific community still has a particular writing style and the results are published in specific formats, in a language, which is often difficult for laypersons to understand (Rakedzon et al., 2017; Shulman et al., 2020). Also, the ever-growing volume of scientific information is challenging to keep track of. In addition, although science communication has become more popular among scientists, presenting scientific evidence to a broader audience is often still an add-on or outsourced to science journalism and public relations due to lack of time (Besley et al., 2018) or lack of training skills (Hundey et al., 2016; Poliakoff & Webb, 2007). Furthermore, science communicators and journalists also have limited resources and must be selective in communicating findings. This can result in biased media coverage (Schäfer, 2012). Apart from what is communicated and how, in view of conspiracy beliefs, for instance, there is no guarantee that the scientific evidence will be taken up and accepted. In this regard, a recent study could show that the persuasive power of LLMs led to a sustained reduction in belief in the conspiracy theory chosen by the participants (Costello et al., 2024). Although there does not seem to be widespread mistrust in science as an institution (Cologna et al., 2025), even small anti-science groups can cause significant damage (Jolley & Douglas, 2014).

Therefore, on the part of communicators, it is obvious to use AI's potential to support science communication to facilitate access to scientific information, prepare this information to target diverse audiences, or boost the production of science communication content, for instance. Thus, as both a channel and a communication partner in science communication (Guzman & Lewis, 2020), AI is about to take over tasks, that were once the domain of science journalists and experienced communicators. Unfortunately, there is a significant lack of research on the potential offered by the combination of AI and science communication (Schäfer, 2023).

At the same time, with the introduction of LLMs, these possibilities for using GenAI in science communicative contexts have also been made widely accessible to laypeople and the general public. GenAI-powered tools offer an answer to every possible request—and disrupt the idea of the availability of pre-tested and exclusively highly reliable sources of information for scientific findings. Indeed, recent surveys have shown that people use GenAI to retrieve scientific information (Greussing et al., 2025) and to answer factual questions (Fletcher & Nielsen, 2024). However, there are significant problems with this technology in meeting scientific standards, and the development of guidelines for utilizing GenAI is cautious (Blau et al., 2024; H. Wang et al., 2024). Thus, users of these tools may be exposed to biased, incomplete, or even misleading information, with a highly persuasive or aesthetic linguistic presentation at the same time (Gottschling, 2024). Despite efforts to make these models more reliable, it must be clear that this contradicts the basic training paradigm of LLMs, which achieve their impressive performance precisely because of the vast amount of data that has neither been checked for accuracy, nor is it grounded in any

model of the world (Bender et al., 2021) with its fullness and complexity.

Nevertheless, GenAI will be used by people for a wide range of queries. Consequently, we face a tool capable of providing answers at a new level of complexity and persuasion to these concerns. Ideally, the objective should be to enable science communicators, researchers, and laypersons alike to make evidence-based decisions on whether to use GenAI as an author of science-focused texts or considering it as a reliable source of information. Therefore, to ensure safe and effective interactions and allow people and the science communication field to benefit while minimizing risks, it is necessary to conduct research that examines the perception of AI authorship in scientific content.

1.3 Determinants for the Acceptance of AI-generated Texts

Before ChatGPT, providing scientific information through AI was strongly linked to the accessibility of structured, machine-readable data (Graefe et al., 2016). Nevertheless, there were attempts at using NLP approaches to make scientific information more accessible. Some examples among many include the Open Research Knowledge Graph (ORKG)[3], which aims to transform scholarly knowledge in a machine-actionable manner, Iris.ai[4], developed to assist researchers in navigating and understanding scientific literature, or SciNote[5], an electronic lab notebook helping researchers manage their data and automatize research reports. For good reasons, these approaches prioritized high scientific standards and reliable data, but were inferior to today's LLMs in application breadth and textual quality. The potential to efficiently and quickly produce news stories in mass was seen early on in journalism, where automation of short and data-driven text pieces was somewhat easier to apply. Thus, research on the perception of actual AI-generated content comes primarily from automated journalism, which focuses on comparing AI authors with human authors.

Two meta-analyses are to be highlighted (for a short overview, see Manuscript I; Lermann Henestrosa et al., 2023), the earlier one comparing 12 experimental studies on news stories and the more recent one considering 30 studies with broader topics. Graefe and Bohlken (2020) found no differences in the perceived credibility of AI vs. human-written articles. Still, higher credibility ratings occurred when people were told that a human was the author. S. Wang and Huang (2024) revealed a small negative effect of AI authorship on credibility perceptions, particularly on sociopolitical and environmental topics. As mentioned, those publications considered primarily data- and fact-driven topics, which were relatively simple in structure and left little room for much interpretation or contextualization. With the expanding possibilities for scientific content, the question arises of how

[3]https://orkg.org/ [accessed April 1, 2025]
[4]https://iris.ai/ [accessed April 1, 2025]
[5]https://www.scinote.net/ [accessed April 1, 2025]

people perceive complex AI-generated texts, particularly on topics that demand detailed explanations of evidence, contextualization, and interpretation of scientific findings.

Due to the relatively surprising introduction of skillful LLMs, research into the perception of their output is still in its early stages. Nonetheless, preliminary work indicates how people could perceive AI-generated output and which factors might play a role. Viewing LLMs through those frameworks can help explain what factors LLMs bring to the table that may promote reliance on their performance.

One of these frameworks is the MAIN Model (Sundar, 2008). It examines how different technological affordances, namely modality, agency, interactivity, and navigability, influence user engagement with online information, which in turn can trigger various heuristics. Thus, it can help in the first place to evaluate a technology based on its sheer appearance and less on the content conveyed.

At first, regarding modality, ChatGPT operated primarily via text, but the integration of multimodality is already being considered, which could enhance its complexity. Audiovisual content could be more effective for certain types of information, for instance, while textual content might be more appropriate for others.

The agency affordance refers to the perceived source of the information. While Sundar (2008) focused on the users' ability to also act as a source of information through interaction, the blurring of who or what is producing the content gets more interesting with LLMs. At first glance, the LLM could seem to be the creator of the output. However, the diffusion of responsibilities poses a problem, especially in the case of sensible information. Furthermore, it is questionable whether something like an expert heuristic can work (e.g., when LLMs cite sources), whether the LLM becomes an expert (e.g., due to its answer on almost any conceivable question), or whether these evaluations are again outsourced from the interaction with the LLM (e.g., by switching to traditional search engines).

One of the most significant strengths of LLMs lies in the almost complete implementation of the interactivity affordance: The fast responses, the adaptability in the conversation flow, and the context consideration make LLMs such as ChatGPT the perfect interaction partner. Interactivity, that is, for example, repeatedly asking questions or getting sensible sounding answers almost independent of the formulation of the question with content of any sort, was not possible before. Especially in the case of scientific information, receiving targeted explanations of scientific topics and avoiding technical jargon could be facilitated through GenAI and help users get actively involved in the debate with science.

While current LLMs provide different display formats to make information easier to capture (e.g., bullet points, headings), integrating real-time search or hyperlinks will further improve the navigability affordance. In the case of recent LLMs, novelty or popularity heuristics are certainly also effective. What is clear is that, as Sundar (2008) emphasizes, „all these affordances have the power to amplify or diminish content effects on credibility

because they indeed deliver the user to the content and could play this role of a moderator in a variety of psychologically distinct ways" (p. 92).

Beyond the MAIN model, one heuristic that can be triggered in human-AI interaction is the machine heuristic. Sundar and Kim (2019) describe it as a "mental shortcut wherein we attribute machines characteristics or machine-like operations when making judgments about the outcome of an interaction" (p. 2). As already indicated in the MAIN model, AI authorship as a source cue, AI-generated content, or an interaction with an LLM can trigger such heuristics, which may be associated with certain expectations and perceptions regarding the interaction or the output (Yang & Sundar, 2024, for an overview, see). Those beliefs can include unbiasedness, neutrality, objectivity, and accuracy of a machine compared to a human (Cloudy et al., 2022; Waddell, 2019). To the same extent that heuristics can generally lead to poorer decisions (Kahneman, 2012), in case of the machine heuristic, it can lead to different behaviors toward an AI, such as greater entrustment of personal information towards an agent (Sundar & Kim, 2019), or the perception that bias in news was attenuated when uncivil comments were moderated by a machine (S. Wang, 2021). Currently, individuals are most probably still in the familiarization phase with GenAI, and there are different forms in which and how transparently it can be built into applications. Moreover, ideas about it can vary with individual prior knowledge and experience. Thus, the extent to which the machine heuristic and other heuristics are activated in interactions with GenAI remains to be investigated.

Similar to the MAIN Model, the Technology Acceptance Model (TAM; Davis, 1989; V. Venkatesh & Davis, 2000) provides a framework to explain which factors lead users to adopt a new technology. As its first version was oriented toward technologies in the work context, Davis (1989) focused on two factors: perceived usefulness and perceived ease of use. Since then, the model has been extended and refined over the years to investigate the determinants of perceived ease of use, to consider subjective norms or also output quality (V. Venkatesh & Davis, 2000), and to combine the determinants to a unified theory for intention and actual usage (UTAUT; Venkatesh et al., 2003). GenAI already fulfills key criteria as it is largely freely accessible, integrated into numerous applications, and not only easy to use, but even superior to traditional search engines and chatbots regarding user-friendliness.

While those prerequisites already increase both perceived usefulness and ease of use—key factors for a broad acceptance—both the MAIN model and the TAM focus on the path to utilization, while in the case of GenAI, the subsequent interaction and the output are the novel and superior features. The content generated is often diverse, unpredictable, and context-dependent, meaning that traditional usage models fall short here. How users evaluate and interpret these outputs and integrate them into their prior knowledge and attitudes and how their expectations and attitudes generally form toward this technology will be decisive for acceptance and is still to be explored.

As an innovation that has rather abruptly become the focus of public attention, research on peoples' perceptions and their basic attitudes toward GenAI is happening almost simultaneously to its introduction. While general attitudes toward AI and specific applications are well-researched, including, for instance, the influence of personality factors (Park & Woo, 2022), the role of the task the AI takes over (Potinteu et al., 2023; Schepman & Rodway, 2020), the perceived capabilities of the AI (Glikson & Woolley, 2020), or the consideration of opportunities and risks (Cave et al., 2019; Schwesig et al., 2023), specific attitudes toward GenAI and ATG are still being explored. For example, Zhang and Dafoe (2019) showed that Americans generally support developing AI applications, but also hold several concerns that should be countered with regulations and safety measures. Such results indicate that a differentiated picture could also emerge in the case of attitudes toward LLMs.

In addition to understanding how such unique technology is now perceived, it is also necessary to assess how actual knowledge about this area is linked to possible concerns or hopes. After all, false hopes and unfounded concerns can harm interaction with LLM-based tools and prevent people from benefiting from them where they have potential.

One field of research that is exceptionally committed to investigating false reluctance to technologies that are superior to humans is that of algorithm aversion. Algorithm aversion is widely described as the reluctance to use superior but imperfect algorithms (Dietvorst et al., 2015) and is investigated primarily in the field of decision making, where the algorithms used usually perform better than humans. However, LLMs exhibit a complexity that does not allow for simple comparisons to where they outperform humans and where not. The processes of searching, receiving, and processing information with LLMs involve multifaceted interactions that must be seen as complex decision situations (for the psychological determinants of LLM-assisted decision making, see Eigner & Händler, 2024), for which users would have to be quite practiced to exploit them to their advantage.

Nevertheless, algorithm aversion research offers insights into scenarios where the rejection of GenAI and its output is more likely. In their systematic review, Burton et al. (2020) specify experience with the algorithm as positively associated with its utilization and expertise in the domain—in the case of human-LLM interaction, the topic of conversation—as negatively associated with it. Jussupow et al. (2020) identify, for instance, perceived capabilities as a factor determining whether users develop algorithm aversion. Accordingly, higher perceived capabilities might increase algorithm appreciation, a factor where GenAI has a head start. Another influence is perceived control over the AI to serve a feeling of user autonomy (Benke et al., 2022) or confidence and suggestibility (Burton et al., 2020; Dietvorst et al., 2018). This can be the users' control in the interaction process—which the chat character and the prompting of LLMs already powerfully serve—but also the presence of a human co-author, especially in the context of ready-presented scientific communicative content. It remains to be investigated, whether it is

solely the presence of an additional control instance that increases trust in AI, whether it necessarily must be a human, how different degrees of co-authorship affect this trust, and what concepts people have of human-AI collaboration.

Apart from external and personal factors, content-related characteristics of AI-generated texts and their appearance to be "too human-like" can also contribute to rejection. Anthropomorphism, which describes the extent to which people attribute human characteristics and behaviors to non-human objects (Epley et al., 2007), can play a role here. On the one hand, perceived anthropomorphism of an agent has been shown to create an emotional connectedness to it (Araujo, 2018); on the other hand, it can also lead to disappointment if expectations are not met in the interaction (Grimes et al., 2021) or the threshold of acceptable human-likeness is exceeded (Ciechanowski et al., 2019; Mori et al., 2012). Furthermore, LLMs may be perceived as less suitable for specific topics or types of texts. Studies examining contexts of rejection show that aversion is more probable when the task is perceived as subjective, requiring intuition, affect, or empathy (Castelo et al., 2019; Huang & Rust, 2018; Longoni & Cian, 2022). This is attributed to the belief that AI systems can neither understand nuances in human language (Graefe et al., 2018), nor provide subjective judgments (M. K. Lee, 2018). Tandoc et al. (2020) specifically compared subjectively and objectively written texts and found a decline in the credibility of AI texts when they were subjectively written. As LLMs are now able to generate texts on virtually any topic in any style, incorporating emotions, interpreting information, and categorizing it, this research must be expanded.

Despite the outlined existing theoretical frameworks on determinants for technology usage, processes during the interaction with communicating machines, and factors influencing aversion or adoption of AI systems, there remains a significant gap in investigating AI as a novel authorship cue on complex text types as well as on people's attitudes and perceptions towards a newly risen technology. Therefore, the present work is dedicated to the investigation of this new source cue and the effect of information on AI authorship perception.

2 Objective

The overarching purpose of this thesis is to systematically explore peoples' perceptions of AI authorship in the context of science communication. Although public awareness of AI-produced content has increased significantly since the launch of LLMs, it is not a novel concept. ATG has been particularly relevant, for example, in automated journalism for several years, where automatic short news reporting was used early. However, with the breakthrough of LLMs, the range of topics for AI to write about has expanded to almost all areas. In addition, the generation of fluent language is now possible in various styles and tonalities, particularly in a remarkably human-like fashion.

These advancements, combined with the increasing availability of scientific data, hold significant potential for science communication, science journalism, and scientific information retrieval. More precisely, the processing, accessibility, interaction, and communication of scientific content could be revolutionized with GenAI as a newly manifesting source. However, research on the perception of AI authorship, especially on more complex topics like scientific content, is still in its early stages. Also, people's attitudes and ideas toward ATG are relatively unexplored. Thus, scientists contemplating whether to use GenAI to facilitate science communication, as well as laypeople considering consulting an LLM for information retrieval, lack the means to adequately assess the benefits and downsides of doing so.

Therefore, this dissertation aims at contributing to a better understanding of the perception of GenAI communicating scientific content by systematically examining the perception of AI authorship (who the alleged author of a text is), the effect of information presentation (how the information is presented), and the influence of disclaimers (which information is provided about AI).

Across a series of controlled experiments, AI-written science communicative articles were presented to investigate how credible and trustworthy readers perceive the texts and their authors to be. Moreover, the presentation of information about the AI served to assess the impact of participant's education via disclaimers on their perceptions. Furthermore, two surveys with representative German samples investigated attitudes and concepts about ATG before and after the release of LLMs and thus complemented the experimental data with current survey data. In doing so, this dissertation provides initial answers on how people perceive AI authorship in more complex content such as science communication, what they think about AI authorship, and what factors may influence this perception.

3 The Present Work

This dissertation is based on several experimental and survey studies, which were combined into three packages. Each package deals with a separate question that serves to answer the overarching research question of how people perceive AI authorship and AI-generated texts in the context of science communication. Before discussing this dissertation's strengths, limitations, and implications, the three Study Packages are summarized below with their respective main findings.

3.1 Evaluative AI: Credibility and Trust Perceptions in Science Journalism

Study Package 1 comprises four online experiments investigating the influence of authorship (AI vs. human journalist) and information presentation (neutral vs. evaluative) on the credibility and trustworthiness of science journalism texts and the evaluation of intelligence and anthropomorphism regarding the respective author. All four experiments were conducted before the publication of LLMs and used the same authorship manipulation. Participants were informed that either an AI had authored the article using specific sources, or that a human journalist had composed the text, utilizing the same specific sources. Both authors, as well as the allegedly used sources, were presented before participants read the article. In addition, a byline below the article and a small picture also drew attention to the respective author. In Studies 1.1-1.3 (Manuscript I; Lermann Henestrosa et al., 2023), the same human-written texts were used[6], while the material for Study 1.4 (Manuscript II; Lermann Henestrosa & Kimmerle, 2024b) was created with GPT-3[7]. Across all four studies, the spread of the wolf population in Germany served as the scientific communicative content.

A basic assumption was that AI authorship should lead to lower credibility and trustworthiness perceptions of the science journalistic text, but only when using evaluative language with emotional word choice. This assumption was based on findings on the machine heuristic and studies on algorithm aversion, which indicate that machines are perceived as being neutral and objective (Sundar & Kim, 2019; Yang & Sundar, 2024) and rejected in tasks that require empathy and emotion (Luo et al., 2019).

The results of Studies 1.1-1.3 show descriptively small but non-significant effects of authorship on message credibility and trustworthiness in favor of the human author. In

[6]The corresponding material is available in the online publication's attachments

[7]The corresponding material can be found in appendix B.2, Manuscript II; data and analysis are available at https://osf.io/gpkc6/

addition, the human author was perceived to be more intelligent and anthropomorphic than the AI. However, evaluative writing also increased the perceived anthropomorphism of the AI author, indicating that the implementation of value-laden language can change the attribution of characteristics and the conception of an AI author. However, the evaluative presentation of information, that is, the presence of evaluative and emotional words in the text, led to overall lower credibility and trustworthiness ratings independent of alleged authorship. In all three studies, the evaluative written text was formulated in favor of the spread of wolves, which aligned with participants' prior attitudes towards wolves. Therefore, a third additional evaluative-negative condition was implemented in Study 1.2 to investigate participants' reactions toward attitude-contradictory material. Again, only a main effect of information presentation occurred with the lowest ratings in the evaluative-negative condition. Thus, the assumption that an AI as author in combination with evaluative information presentation leads to different reactions compared to a human author, that is, more negative perceptions, was not confirmed. An assessment of the machine heuristic in Study 1.3 resulted in a stronger attribution of the characteristics of accuracy and objectivity, for instance, to an AI than to a human journalist. This difference was even more pronounced after the participants had read the supposedly AI-written article. Together with the relatively positive general attitudes toward AI, the presence of a machine heuristic, further activated through the presented text, could explain the high credibility and trustworthiness ratings.

To replicate these findings with actual AI-written material and test equivalence of human and AI authorship with a sufficiently large sample size, Study 1.4 focused on the authorship comparison with only neutral information presentation. An equivalence test (Lakens et al., 2018) revealed a small but significant effect of authorship on message credibility perceptions in favor of the human author. The results on perceived intelligence and anthropomorphism were replicated. Moreover, an exploratory analysis of source credibility and behavioral intentions revealed significantly higher ratings for both variables when the alleged author was a human. In sum, this package suggests a significant but small difference between a human and an AI author such that the human author was perceived to be more credible. However, it was primarily the information presentation that influenced the perceptions of the texts regarding all dependent variables.

At the time of the studies, explicitly declared AI-generated content was still rarely distributed. Despite the fact that there seem to be small differences in favor of a human author, the high credibility and trustworthiness ratings for the AI in the neutral condition speak for a fundamental acceptance of AI authorship, even in more complex and longer articles compared to the text types researched before. This picture is also supported by the positive general attitudes toward AI in Studies 1.1 and 1.2 and the positive machine heuristic in Study 1.3. What remained open was the question of what participants imagined AI authorship to be and on which specific concepts this perception is based.

3.2 Attitudes toward and Perceptions of Automated Text Generation

Study Package 1 provided evidence of only slightly lower credibility and trustworthiness ratings for AI authors compared to human authors. In all experiments, however, the authorship manipulation contained a short information about the functioning of the AI and the sources used. Hence, the conception about it was not completely left to participants. The fundamental question of what people generally imagine AI authorship to be, what concepts they have about it, and their attitudes toward this technology remained open.

Therefore, Study Package 2 systematically recorded attitudes, knowledge, and concepts on AI and ATG. This package comprises two representative online surveys (Manuscript III; Lermann Henestrosa & Kimmerle, 2024a) and a data descriptor (Manuscript IV; Lermann Henestrosa & Kimmerle, 2024c) which includes the two data sets and prepares them for further use[8]. Attitudes and concepts toward AI in general and specifically in relation to ATG were surveyed in the German population before (March 2022) and after the release of ChatGPT (July 2023). The samples were representative regarding age, gender, and educational level.

The results showed that, on average, individuals held relatively balanced attitudes toward AI and ATG. At the same time, there was significant uncertainty regarding aspects such as responsibility for the content, data sources, the role of humans in the generation process, and the creation process of AI-generated texts. When given several options for the respective concepts, around a third of respondents consistently chose "perhaps". This did not change in the second survey eight months after the release of ChatGPT. Unsurprisingly, the general population often does not have a clear idea of how new technologies work or where precisely the content comes from. Besides, it is highly dependent on the specific technology and the way it is implemented in applications. Either way, knowledge about the underlying technology is not a necessary requirement for usage. However, a lack of clarity about responsibilities or opacity about the output sources can be problematic when using LLMs to retrieve scientific information and generate sensitive content. For example, the data basis for very established research fields may be sufficient to produce reliable information. However, there is a risk of misinformation, particularly in young research fields or areas where there is no scientific consensus. Currently, it is at least questionable whether uncertainties or the tentative nature of scientific findings can be reliably reproduced by AI. In addition, the results of the knowledge test in Survey 2 indicate misconceptions about the actual capabilities of AI, for instance, regarding the inability of AI to understand language like a human or not always giving correct answers.

The comparison between the surveys shows only minor differences in attitudes toward

[8]Data and analysis are available at https://osf.io/sn75h/

AI in general and specifically toward ATG at the two time points, with general attitudes toward AI and the belief in the machine heuristic slightly decreasing. Multiple regression analysis with k-fold cross-validation (Géron, 2023) revealed that specific attitudes toward using ATG and general attitudes toward AI were positively related to the intention to consume AI-written texts, while prior experience and knowledge about ATG were not. Looking at the two groups of people who had already used ChatGPT compared to those who had not, the users stand out with more positive attitudes toward the technology, a slightly more pronounced positive machine heuristic, and higher self-assessed knowledge of AI. When asked for a direct comparison between human and AI authorship on 18 topics typically covered by news media, participants indicated a clear preference for the human author, with constantly a third stating no preference for either author. Furthermore, individual item analyses showed that in Survey 2, 72.45% of respondents were in favor of labeling requirements for AI-generated texts, 58.05% agreed that policymakers should set rules for the application of ATG, and 34.06% were in favor of ChatGPT being used also for scientific information (40.08% being undecided on this question).

Overall, the results suggest that around the advent of LLMs, relatively positive attitudes existed toward this technology, which were, however, based on little objective knowledge about AI and ATG. Moreover, the relationships between attitudes and intention to use indicate possible conditions that could promote or hinder the use of new technologies, although the exact direction is still unclear. Even though these are two snapshots in a very dynamically developing field, the need to address the capabilities and limitations of generative AI becomes clear to contribute to a more realistic picture within the population and informed use. Compared to Study Package 1, in which perceptions of AI authorship on a science communication topic only slightly decreased, the results of the direct comparison point toward a pronounced preference for human authors. Together with the more positive attitudes of people who already had experience with ChatGPT, this suggests that the evaluation of AI output in direct interaction could be more favorable than the hypothetical question about preference.

3.3 The Effects of Disclaimers on Readers' Perceptions of AI Output

Due to the uncertainty and lack of knowledge about the capabilities of text-generating AI revealed by the surveys in Study Package 2 and the experiments of Study Package 1 showing reliance on AI output even on complex text types, Study Package 3 investigated the influence of prior information on evaluating AI-generated texts. Providing information in advance is an option for addressing misconceptions and making interactions with AI tools more informed and safer. This is often already practiced via disclaimers to exclude

liability. Here, for example, specific uses can be generally ruled out, or information on the limitations of a tool can be provided to the user during or before the interaction.

Therefore, in three online experiments (Manuscript V; Lermann Henestrosa & Kimmerle, 2025), information about ATG and LLMs was introduced as short text pieces presented directly before an AI-generated text. In case of information about strengths, it included details such as the ability of LLMs to rapidly produce high-quality text, while in the case of information on limitations the lack of factual accuracy was addressed, for instance.

In Study 3.1, participants were provided with information on the strengths, the limitations, or only basic information about ATG before being supplied with a short AI-written article on climate change. Moreover, as a within-subjects factor, information presentation (neutral vs. evaluative) was varied to investigate the interaction between prior information and writing style on the perception of AI-written texts. As in Study Package 1, evaluative information presentation, that is, adding emotional wording to the article, decreased perceived credibility and intelligence but increased anthropomorphism perceptions. While no differences of the disclaimer manipulations were found concerning message credibility, a significant effect of the strengths condition compared to the other conditions occurred for source credibility. An exploratory analysis of the machine heuristic revealed higher ratings for the AI author than for a human author on this variable similar to the results of Study 1.3.

In Study 3.2, again prior information was presented to participants before reading an article about fat-shaming. In contrast to Study 3.1, the information presentation condition was dropped, and the three different disclaimer types included information about the limitations of AI, about limitations of AI and humans, or, again, only basic information. Moreover, as authorship continues to move toward different degrees of collaboration between humans and AI, a human-AI co-authorship condition was introduced, stating that the AI-written article was checked and revised by a human. Thus, I aimed to assess the interaction of information on AI and human flaws with varying authorship attributions. No main effects on message credibility were found. Regarding source credibility, significant interactions suggest more negative evaluations after reading about AI limitations only for sole AI authorship. This points to the potential balancing effects of a human involved as informing about limitations did not decrease source credibility perceptions in the co-authorship condition.

In Study 3.3, additionally a balanced condition, which provided information about both strengths and limitations and a condition without any information were introduced. As in Study 3.2, a co-authorship condition presented between subjects was implemented, this time with an emphasis on the aspect that a human has consulted an AI for help. Like in Study Package 1, credibility, intelligence, anthropomorphism perceptions, and behavioral intentions were assessed. Again, no effects on perceived message credibility

occurred. However, regarding source credibility, the strengths disclaimer produced a significant effect compared to the other disclaimer types. Interestingly, comparing the limitations condition with the other conditions was significant, too, but with higher ratings for the limitations disclaimer on source credibility. Thus, source credibility increased in both the strengths and the limitations conditions. Moreover, the co-authorship was perceived to be more anthropomorphic than the sole AI authorship. As in Study Package 1, this shows that labeled authorship can have a significant impact on the perception of such properties. However, no effects on perceived intelligence and behavioral intentions occurred.

Across the three experiments, only small effects of disclaimers on source credibility were found, arguing against a large influence of prior information on evaluating subsequent material and AI authorship. However, high exclusion numbers due to failed manipulation checks indicate a lack of clarity in the information presented and an inability to reproduce the disclaimer information. The question arises if attempts to inform users about potential pitfalls of a technology, for example, in the interface of current LLMs, fulfill the intended purpose.

4 General Discussion

This dissertation combines two topics—science communication and GenAI—not just for their individual importance but because precisely this combination gives rise to several socially relevant issues beyond the particular subject areas. Science as an institution, its results, and the scientists who produce them face threats from political interests (Forchtner et al., 2018), nationalist and populist tendencies (Mede & Schäfer, 2020), conspiracy theories (Connolly et al., 2019; Harambam & Aupers, 2015) and multiple global crises (Lasser et al., 2020). In addition, the growing wealth of information (Bornmann et al., 2021) makes it increasingly difficult for interested laypeople, science journalists, and scientists alike to gain an overview of scientific findings and adequately assess their quality and validity. Here, AI holds great potential to address the challenges that arise from the abundance of information, to support scientific work, and to improve the communication of scientific content. However, exceptionally high demands are rightly placed on the scientific knowledge process and its results, which should also be met by technologies that are intended to support this area. In the long term, rules and ethical guidelines are essential for safely using AI-based tools in scientific contexts but have not yet been sufficiently developed. Developing such guidelines requires knowledge of how people use these tools, what they expect of them, and how they perceive them to ensure that the guidelines answer relevant user needs. The present dissertation contributes a first step in this direction by uncovering the opportunities that AI offers in science communication and the limitations that urgently need to be addressed.

As the presented studies showed, AI, as an important tool to produce and disseminate information, including science communicative content, seems to be trusted by many people. Unexpectedly, in Study Package 1, a transparently declared AI author on a science communication topic that primarily concerns humans and on which AI could not have communicated until the time of the first studies was not rated much more negatively than an experienced human science journalist. Accordingly, following participants' demand to label AI-generated texts which they voiced in Study Package 2, should not be linked to expectations of a change in the evaluation of labeled AI texts. What has been shown overall is that the evaluation of a text depends on the presentation of the information. This suggests that text characteristics may play a more important role than authorship attributions, even if the author lacks real understanding, human intelligence, and empathy. Exploratory analyses in Study Package 2 also indicated that attitudes toward AI in general and ATG are significant predictors of behavioral intentions. These predictors for the adaptation of technologies, which have already been studied for a long time, should, therefore, be investigated in future studies in the context of science communication. At the same time, in Study Package 3, prior information did not significantly impact cred-

ibility perceptions, suggesting that attitudes may be somewhat robust and resistant to change through simple disclaimers.

4.1 Merits

Technological development is happening rapidly and is significantly impacting people's everyday lives. However, consideration of the psychological impact of introducing such innovations often falls by the wayside. Research into the perception of AI authorship prior to the publication of ChatGPT was conducted almost exclusively in the field of automated journalism and this area of research was strongly characterized by studies on the perspective of potential job losses in journalism (e.g., Carlson, 2015; Kim & Kim, 2018; Latar, 2018). Therefore, the perspective on other areas such as science communication but also on broader user perceptions is lacking. Hence, this dissertation contributed to closing a critical gap, as scientific research risks falling behind technological advancements. It bridges various disciplines, such as science communication, automated journalism research, and psychological human-AI interaction studies, which must necessarily converge to examine the effects of AI technologies on human perception. Specifically, it extends automated journalism research by linking it with psychological studies on trust and credibility to address urgent questions arising from current developments in AI.

Thus, as one of the first studies, the research presented here dealt with the potential that was to be anticipated from developments in the field of NLP even before the public attention on LLMs. As to my knowledge, this work is one of the first to combine upcoming developments in GenAI with their possibilities for science communication. Moreover, by using topics typically covered by science communication not only the capabilities of current LLMs were appropriately considered early on but also the possibility was offered to investigate this novel authorship cue in a field where its limitations can be particularly critical.

It should, therefore, be understood as accompanying psychological research on current technological developments in the practical application area of AI. With their variety of methods and well-founded theories on the acceptance of new technologies and paradigms in human-AI interaction, psychological studies such as the present work are essential for understanding the effects of new technologies on the perception and acceptance of people (Buder et al., 2024) and form the basis for the positive and ethically sound design of human-AI interactions.

Even though the surveys only represent snapshots, they do offer the possibility of a comparison, especially at two particularly critical times surrounding the public attention on LLMs. The results are unsurprising given the lack of focus on this area in previous years. In addition, the response distributions can also be found in similar proportions

in other surveys (e.g., MeMo:KI, 2022). Nevertheless, they indicate which factors could determine usage and where there could be potential misunderstandings regarding the expectations toward AI. Together with the data descriptor (Manuscript IV), the surveys represent one of the first comprehensive assessments of people's perceptions toward ATG and LLMs, which can serve as a basis for further research and investigating specific questions.

Furthermore, the methodological diversity, which enables different perspectives on the evaluation of AI authorship and current attitudes toward it through experimental designs, survey studies, and various analysis methods (equivalence testing in Study 1.4, k-fold cross validation in the surveys, as well as analyses of variance in Study Package 1 and multiple regression analyses with custom contrasts in Study Package 3), served to shed light on the research questions from different angles. Multiple replications of the authorship effect, the information presentation effect, and the disclaimer effect on various variables further demonstrate the results' robustness.

Overall, all three study packages comprise pioneering studies that address various research questions in a still-young field of research. Although they cannot provide conclusive answers to the perception and acceptance of GenAI in science communication, the initial insights nevertheless offer a valuable starting point for future research.

4.2 Limitations

In the following section, I will examine the limitations of this dissertation and discuss possible reasons for the lack of effect of the interventions in the experimental studies.

The material was initially self-generated as the first experiments were conducted before the availability of high-quality LLMs. Although GPT-3 and later GPT-4 were used for the latest experiments, the studies were designed to present finished output. Therefore, the scenarios examined here only partially correspond to the current technological and usage possibilities, which now allow for direct interactions. Accordingly, the results on the perception of these texts are most applicable to traditional forms of science communication, for example, on websites, where finished texts are provided for direct reception by the reader. Thus, this work does not cover all GenAI capabilities and application areas.

In relation to this, when content is not communicated directly by an AI but provided via science communication platforms, for example, trustworthiness and credibility perceptions may be higher due to the reputation of the website (Haas & Unkel, 2017). Additionally, it should be considered that the participants might have perceived the study setting and the experimental investigators as an already passed verification instance, which could have weakened AI authorship as a direct communicator and reduced already minor effects. Thus, unmediated interactions with GenAI must be distinguished from those scenarios,

as they leave the verification and validation of the content to the user and are, therefore, more critical.

Especially in science journalism, which is usually characterized by a careful selection of topics, thorough research, and differentiated positioning in the context and life situation of the people, it is questionable if AI can reliably curate this field with no human in the loop. Therefore, the sole AI authorship suggested in the studies of Study Package 1 could be a relict to which the investigation of human-AI co-authorship in Study Package 3 was an attempt to respond. However, as indicated by the exclusions due to failed manipulation checks regarding co-authorship, there may be different interpretations as to which contributions lie with the human and which with the AI and who is, therefore, to be identified as "the author". This was not systematically checked or queried in Study Package 3 but is a challenge future research could tackle.

The selection of topics can be viewed critically from two perspectives: first, regarding the subjects' attitudes toward the respective topics, and second, because the topics were predetermined. The results of the studies in which the subjects' attitudes to the topic of the text were recorded show that participants tended to have attitudes that corresponded with the content of the article. However, the findings from Study 1.2 indicate that attitude conformity alone cannot account for the absence of authorship effects, as no authorship effect was observed even with an attitude-incongruent text. Nevertheless, future studies should take into account personal relevance, involvement, or more substantial individual interest. This is supported by the results of the surveys, which show a preference for the human author in a direct comparison and indicate variance between the topics. Especially when topics are researched out of personal motivation, the effects of the fit with opinion and expectation could be stronger. However, many topics in science communication do not affect individuals directly at the moment of communication.

Apart from the content, which was perceived as positive by the participants, the articles were error-free. On the one hand, those conditions left little room for doubt about the information and might have caused a positive expectation disconfirmation, as open comments of the participants sporadically suggest (e.g., "I found it surprising how naturally the texts were written", "I found the article very interesting and well written", "I am impressed by how far AI systems have come"); on the other hand, this reflects actual technologies' capabilities and the providers' efforts to meet their users with answers that appeal to them and are tailored to their needs. Thus, the scenarios used are highly realistic. It is no longer appropriate to introduce errors at a linguistic level. However, whether errors in content, unless they are obvious, can be detected by participants during the experiment, is open to doubt. With regard to the presented studies, the participants had little reason not to believe the content, even if it came from an AI.

Overall, the experimental conditions may have contributed to creating an optimal setting for the participants when confronted with AI-written texts, which can be seen

as a limitation and an additional reason for the small or absent effects. Especially in Study Package 1, when ATG was still hardly known, the sources the AI had allegedly drawn were stated, besides the information that the article was allegedly published in a newspaper. This was done with the intention of creating a credible cover story and keeping the conditions as constant as possible for both authors. Hence, this transparency about the sources the AI had allegedly used and the high linguistic quality of the articles, reflecting most participants' opinions on the topic, left little room for questioning or rejecting the material and authorship.

This is also related to the selection of the dependent variables, that warrants critical discussion with regard to their interpretability in terms of acceptance. Credibility and trustworthiness are central conditions for high acceptance of texts (Sundar, 2008), predictors of behavior intentions, and use behavior (Kelly et al., 2023). Due to the large number of credibility measures, we chose the Message Credibility Scale (Appelman & Sundar, 2016), specifically developed for news, as it focuses on assessing the credibility of the specific content within the article. In contrast, the Trust in News Media Scale (Kohring & Matthes, 2007) seeks to capture a broader, more reflective assessment of the article and the journalistic work as a whole. However, a reanalysis of the data of Study 1.3 revealed a strong positive correlation between message credibility and trustworthiness (r = .74). This suggests a substantial overlap between the two constructs, which potentially measured similar aspects and cannot be seen as distinct. Similarly, a reanalysis of the data of Study 3.3 revealed a strong correlation between message and source credibility (r = .67). This is not surprising, as the scales used were the same and differed only in the formulation of who they were aimed at (text vs. author). Nevertheless, future studies should, if possible, strive for measures that better distinguish the credibility of the text purely in terms of content and of the stated author in general.

Moreover, as measured in the presented studies, the two constructs primarily reflect snapshots of specific, predetermined outputs. Accordingly, it is quite possible that a ready-presented text is rated as relatively credible, but the basic acceptance of AI-generated content varies. Other important questions to capture acceptance could have been about how much participants liked an AI writing about the topic or how they felt about an AI expressing an opinion. By querying behavioral intentions for further usage of AI-generated texts, an attempt was made to come closer to a more general acceptance at the behavioral level. However, the actual use, including the frequency, quality of interaction, and type, must be the subject of longitudinal studies to draw conclusions regarding sustainable acceptance and its determinants. Given that GenAI enables extensive interaction with digital content far beyond the traditional consumption of scientific information, additional variables related to the interaction experience should be considered. These can include verification strategies following an interaction or content engagement and processing depth.

Finally, the results of this dissertation must be considered in a contemporary context. ATG is a rapidly evolving field that has gained tremendous public attention due to recent developments. Most of the studies included here were conducted before this new technology became the focus of attention. However, public debates can strongly influence perceptions, attitudes, and usage behavior (Anania et al., 2018; C.-J. Lee & Scheufele, 2006; McCombs, 2004). In addition to that, all studies (except Study 3.2) examined German-speaking samples. Although little research has been conducted on intercultural differences, these could influence the effects and attitudes toward AI (Ge et al., 2024; D. Ma et al., 2024). Thus, generalizability to other countries in which the attention and debate around GenAI might look different is limited. Therefore, more research—especially longitudinal and cross-cultural work—is necessary to further validate the results found here.

4.3 Theoretical Implications and Future Research

The release of LLMs as highly user-friendly, interactive, and responsive chatbots with intuitive interfaces demonstrates the developers' success in effectively implementing and integrating the criteria outlined in the MAIN model (Sundar, 2008) and the TAM (Venkatesh et al., 2003). The increasing integration of new features such as real-time search, customization for specific needs, and enhancements in multimodality represent further advancements of already highly user-friendly and accessible GenAI applications. While these frameworks have been effective in explaining the varying degrees of adaptation of applications, they fall short in explaining usage patterns, user interactions, and users' fundamental perceptions of the generated output. The focus should now shift to researching the heuristics that can be activated during the interactions with AI-based systems and understanding the implications of their activation. Future research could benefit from integrating the heuristics outlined in the MAIN Model with the attitudes toward using the technology provided in the developments of the TAM. This would enable a more comprehensive examination of the conditions users bring to their interactions with technology, which can influence their decision to use or not use GenAI, the manner in which they use it, and their uptake and acceptance of the output. As this work has demonstrated, people can attribute distinct characteristics to AI and human authors and vary in their attitudes toward AI and ATG. In the field of science communication, which is focused on specific outputs and interaction scenarios with LLMs, it is crucial to investigate users' perceptions and actual interactions in these specific contexts. Only by this it is possible to assess whether these interactions meet the objectives of science communication and to identify any obstacles that need to be addressed.

With reference to the literature on algorithm aversion and appreciation, the question

arises whether small differences in favor of the human author and preference for human authorship in direct comparisons still speak for AI-aversion. Alternatively, these differences could even be interpreted as appreciation, given the overall high credibility ratings. The survey results indicate that people tend to find AI authorship more suitable for certain domains than others, suggesting a nuanced perception of AI's appropriateness. The results of Study Package 1, however, indicate that when the output is of high quality, people might be willing to accept AI authorship even for issues that significantly impact humans. Considering that the human author in Study Package 1 was an expert in his field and that information on the AI's limitations in Study Package 3 did not lead to a decrease in the evaluation of the AI's text, it is difficult to interpret the minor differences as a rejection of AI authorship for science communicative content. Accordingly, in light of this work, it is not possible to definitely speak of either algorithm aversion or appreciation, but rather an evenly matched scenario. In regard to the growing body of literature failing to find evidence of algorithm aversion (Hou & Jung, 2021; Logg et al., 2019; Mariadassou et al., 2024), it is reasonable to assume that this effect is diminishing. However, scenarios are conceivable in which a preference for human authors could become more pronounced, for example, through the variation of topics. Future research could explore, for instance, whether readers exhibit different preferences when given the option to choose between authorship on specific topics.

In decision-making literature, people can typically compare their judgments with those of an expert or another person versus AI, resulting in different findings depending on the reference (Jussupow et al., 2020). Looking at science communication targeted at the public, it could be argued that any information might be valuable to laypeople, regardless of who communicates it. Thus, relying on an imperfect AI might still be superior for laypeople with limited knowledge. Following this, the results presented here could be seen as a sign of placing trust in an AI that has more information than most participants. However, since GenAI-mediated science communication clearly differs from prediction algorithms that perform better than humans, it requires a more detailed differentiation of when it might be good to rely on GenAI output and when not (S. Ma et al., 2023). Moreover, the studies here cannot answer whether acceptance of AI authorship is based on genuine trust and (over)confidence in AI or if it stems from a lack of confidence in one's knowledge or the perception of oneself as a layperson. The results regarding the machine heuristic suggest that even a human journalist, who is an expert in his field, is considered to be more biased and less objective than an AI. Regardless of the specific domains in which AI surpasses human capabilities, it is essential that future research addresses pre-existing expectations and concepts, as they are likely to influence how individuals apply these expectations to other interactions with AI.

When speaking about over-reliance, the implicit assumption is that sometimes it might be inadequate to trust AI. As the accuracy rates of LLMs in different areas prove (Ope-

nAI et al., 2023), relying on an LLM's output can be a proper decision regarding some topics but not for others, especially when there are consequences tied to this reliance (Buçinca et al., 2021; Eigner & Händler, 2024; Vasconcelos et al., 2023). In the case of science communication to laypeople, detecting "errors" or hallucinations will be even more challenging as they are persuasively communicated and not easy to spot (McGowan et al., 2023). Especially for young research fields, where available data is limited or regarding sensitive health-related information, where high accuracy must be the maxim, some degree of educational information might be appropriate. It could be that people are becoming more and more aware that these systems can be wrong or produce misinformation. Thus, it is to be investigated if people rely less on LLMs answers if they become more cognizant that they can err or even see them producing inaccurate content (Dietvorst et al., 2015; Nourani et al., 2020). However, since it requires a relatively high level of expertise to recognize subtle inconsistencies, exposing false or incorrect answers in the case of non-experts in scientific domains is much more challenging. Particularly in science communication, where laypeople have mostly no choice but to trust as they cannot perform a first-hand evaluation (Bromme & Goldman, 2014), they will have a hard time verifying or disproving information. This responsibility cannot be placed solely in the hands of users. It remains to be seen whether a form of informed decision-making will develop, enabling individuals to discern which specific tasks can be safely delegated to GenAI.

To support this, interventions are to be developed as to how teach people the capability of deciding when it is good to consult an LLM and when to invest in verifying the information it presents. This requires a shift in focus toward understanding the implications of both relying on and disregarding AI-generated output, which consequences it has when AI communicates topics beyond its capabilities, and the potential risks if individuals accept such information without questioning it enough. For sure, people also learn to handle new technologies by doing and get an intuitive feeling of their capabilities and weaknesses. However, this can vary individually and depend on the systems' transparency and understandability (Bansal et al., 2019). For a technology specifically designed to enable pleasant and engaging conversations, additional measures will probably be required to mitigate its inherent disadvantages.

At the same time, the approach pursued in this dissertation emphasizes that detailed technology-specific capabilities are initially of secondary importance when examining the perception of AI-generated texts. On the one hand, the performance of AI depends on rapid development and is, therefore, quickly outdated. On the other hand, it is strongly determined by the requirements of the respective field of application, the restrictions imposed by providers, and the ability of end users to exploit these (Eigner & Händler, 2024; Wei et al., 2022). However, research committed to the human perspective should not focus on manipulations, such as the exhaustion of artificial scenarios, or the artificial manipu-

lation of AI output, such as the introduction of errors. This presents the risk of lagging behind technological development and submitting to it. Nevertheless, this approach does not contradict an in-depth examination of the actual capabilities of GenAI. However, simply testing each new model and focusing on what-if scenarios does not take seriously the capabilities of LLMs and the developers' motivation to answer every conceivable question in various ways, to pursue even more user-friendly technologies, and to implement them in nearly every application possible. Thus, since people will be increasingly confronted with AI-generated content in the future and will also interact with it more frequently, while being more or less aware of it, research is needed both on the perception of the static output and AI authorship, as well as on the dynamic interaction behavior and the processes taking place there. The disciplines of psychology and cognitive science, with their fundamental theories of learning, perception, and information processing, offer an ideal basis for this. Only in this way can we develop a deep and comprehensive understanding of human perception and interaction with AI in the long term that focuses on the well-being of people.

In future studies it will be essential to distinguish current expectations (such as heuristics) from overarching attitudes and content evaluations. In the studies of Study Package 1, it was unclear whether participants were positively surprised and thus rated the text favorably, or if the text itself was simply of high quality and the authorship was of lesser importance, or both. For instance, users might have a positive machine heuristic, believing that the AI is competent in writing about a specific topic, yet find the text unappealing. Conversely, one might approach an LLM with negative expectations or also having general reservations toward LLMs but find the text acceptable. Both scenarios can result in similar text evaluations. Therefore, not only the specific expectations people can have when confronted with AI will be of particular importance (Grimes et al., 2021) but also how they weigh them compared to when engaging with texts written by humans.

The results concerning anthropomorphism and intelligence suggest a differentiation in author perception, which may become significant in contexts where perceived human-like qualities play a more crucial role or in which AI might be perceived as inferior to human judgment. Further investigation into perceived intelligence and anthropomorphism is warranted, as it can elucidate the specific aspects in which users perceive AI to be superior to humans, or whether these perceptions are based on misconceptions about the actual capabilities of AI. Based on my previous discussions, it is essential to note that under the examined conditions, a preliminary acceptance can be observed, which should be further challenged in subsequent studies.

In sum, a systematic assessment of people's prior attitudes, heuristics at work, expectations related to AI regarding specific topics, and the mental models they possess regarding AI functionality could provide deeper insights into the prerequisites and conditions influencing perceptions of AI-generated content. An understanding of these individual

conditions might be essential for evaluating how manipulations of authorship, content, presentation style, or prior information affect the perception and assessment of AI texts. As currently no single theory is sufficient to capture the acceptance and handling of GenAI comprehensively, existing models need to be expanded to include aspects such as trust, control options, and epistemic security in dealing with AI-generated content.

4.4 Practical Implications

The present work has a clear message for researchers, science communicators, and practitioners alike. There is no fundamental aversion to AI-generated texts on science communication content. Still, there may exist general misconceptions about LLMs and AI-generated texts that can thwart a critical and responsible approach to AI output. Addressing those and communicating pitfalls may not be easily possible through simple education. The hope of being able to inform readers economically through presenting information or even to guide them to an informed and safe use appears to be too simplistic an approach considering the results of Study Package 3. It is questionable whether similar forms to disclaimers directing attention to educative information can be fruitful in real, uncontrolled scenarios. Future research should explore to which extent straightforward explanations can help counteract misinformation, especially on topics where LLMs have limited training data. However, the presented findings suggest that disclaimers could primarily serve their original purpose of protecting providers, not users.

Beyond concerns and interventions for the individual, the developers and companies behind GenAI, as well as the providers and distributors of AI-generated content, have an exceptional responsibility. The identification and elimination of vulnerabilities, as well as efforts to make technologies more secure and identify safe application scenarios, should be given high priority. In this regard, Lockey et al. (2021) have identified several challenges faced by various stakeholders int the context of AI and propose a multi-stakeholder approach. Efforts such as fine-tuning, further training rounds with reinforcement learning from human feedback (RLHF), or built-in real-time search are already contributing to this but are primarily post-hoc interventions. Achieving absolute safety, especially in critical and sensitive areas such as health information or insights into emerging research fields, may be challenging due to the nature of LLMs. A debate on where people might be right to reject AI output and algorithmic advice, that is, when accuracy is not given, and consequences of reliance could be critical, is indispensable. Political efforts can provide a legal framework—such as the Digital AI Act, which is binding in the EU—but are difficult to introduce, given the broad incorporation of AI-based tools and their application in almost every conceivable sphere of life. Also, general guidelines can aid orientation but are still to be developed for the individual application areas and probably do not

reliably prevent intentional and unintentional misuse. Either way, without insights into how people actually interact with these technologies and whether or how, for example, a labeling or a disclaimer requirement would have any effect at all, such initiatives might remain incomplete and possibly ineffective.

Besides the described challenges, the acceptance of AI-generated texts by the public should be viewed as a significant opportunity. Specifically in the realm of science communication, AI offers the potential to disseminate complex information and utilize formats that traditionally demanded extensive human resources. Thus, this technological advancement not only facilitates easier implementation of science communication but also is likely to be well-received by the general population. Science communicators can, therefore, be relatively optimistic about projects to promote public interaction with the help of AI. In any case, transparency not only about the applications used but specifically about their limitations should be made a top priority. What may be collectively pursued by science communicators, scientists, journalists, and actors beyond these domains is a critical reflection on the necessity of integrating imperfect technologies into every aspect of science and science communication. In particular, where the limits of GenAI are already known, a heightened sense of responsibility is essential—especially given that AI, by its nature, is incapable of assuming such responsibility.

5 Conclusion

GenAI, especially LLMs, has become one of the most important and disruptive technologies in recent years. It has the potential to fundamentally change the way knowledge is generated, communicated, and processed. At the same time, however, those tools come with specific risks that compel users to adopt a reflective and responsible approach, unlike few developments before. An essential step on this path is to research the human perception of this technology.

This dissertation contributes in several ways by examining the perception of a new type of source cue—AI authorship—while considering the actual capabilities of current LLMs to imitate human written language in an unprecedented manner. Moreover, it sheds light on attitudes and concepts toward text-generating AI at two critical points and examines ways to make the reception of AI-generated content more informed and thus safer. The results indicate that AI-generated content maintains a comparable level of perceived credibility, even for complex topics such as science communication. At the same time, this young technology is still subject to ongoing public discourse, with prevalent misinformation and uncertainties. However, simply providing information before the consumption of AI content seems to be insufficient to influence readers' perception.

The multi-method research approach allowed for the examination of a previously little-studied field from different angles. By conducting several replications, carefully varying experimental manipulations, and using various analysis methods, the robustness of the results was confirmed. Thus, this work represents a first step towards investigating the psychology of human interactions with elaborated AI-generated texts that meet the current standard of LLMs and provides impetus for a critical approach to GenAI. Beyond this work, a deeper investigation of human-AI interaction is necessary. Hereby, an understanding can be developed on how and why people use GenAI, what approaches they take, what challenges may arise, and what outcomes they achieve. Examining actual interaction behavior is the basis for designing these technologies safely and with the greatest possible benefit for people.

As Bender pointed out, we have created machines that mindlessly generate text[9]. But as Burton et al. (2020) stated, "neither blind neglect nor blind acceptance [...] can be considered successful in this view because such decisions signal the absence or failure of interaction between the human and algorithm" (p. 221). The key challenge now is to find ways to enable users to use this technology ethically, responsibly, and for the maximum benefit of society—without ascribing to it more intellect than it possesses and without placing more trust in it than it deserves.

[9] https://nymag.com/intelligencer/article/ai-artificial-intelligence-chatbots-emily-m-bender.html [accessed April 1, 2025]

6 References

Abukmeil, M., Ferrari, S., Genovese, A., Piuri, V., & Scotti, F. (2022). A Survey of Unsupervised Generative Models for Exploratory Data Analysis and Representation Learning. *ACM Computing Surveys*, *54*(5), 1–40. https://doi.org/10.1145/3450963

Anania, E. C., Rice, S., Walters, N. W., Pierce, M., Winter, S. R., & Milner, M. N. (2018). The effects of positive and negative information on consumers' willingness to ride in a driverless vehicle. *Transport Policy*, *72*, 218–224. https://doi.org/10.1016/j.tranpol.2018.04.002

Appelman, A., & Sundar, S. S. (2016). Measuring Message Credibility: Construction and Validation of an Exclusive Scale. *Journalism & Mass Communication Quarterly*, *93*(1), 59–79. https://doi.org/10.1177/1077699015606057

Araujo, T. (2018). Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior*, *85*, 183–189. https://doi.org/10.1016/j.chb.2018.03.051

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019). Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *7*, 2–11. https://doi.org/10.1609/hcomp.v7i1.5285

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

Benke, I., Gnewuch, U., & Maedche, A. (2022). Understanding the impact of control levels over emotion-aware chatbots. *Computers in Human Behavior*, *129*, 107122. https://doi.org/10.1016/j.chb.2021.107122

Besley, J. C., Dudo, A., Yuan, S., & Lawrence, F. (2018). Understanding Scientists' Willingness to Engage. *Science Communication*, *40*(5), 559–590. https://doi.org/10.1177/1075547018786561

Bijker, W. E., Bal, R., & Hendriks, R. (2009). *The Paradox of Scientific Authority: The Role of Scientific Advice in Democracies*. The MIT Press. https://doi.org/10.7551/mitpress/9780262026581.001.0001

Blau, W., Cerf, V. G., Enriquez, J., Francisco, J. S., Gasser, U., Gray, M. L., Greaves, M., Grosz, B. J., Jamieson, K. H., Haug, G. H., Hennessy, J. L., Horvitz, E., Kaiser, D. I., London, A. J., Lovell-Badge, R., McNutt, M. K., Minow, M., Mitchell, T. M., Ness, S., … Witherell, M. (2024). Protecting scientific integrity in an age of

generative AI. *Proceedings of the National Academy of Sciences*, *121*(22), Article e2407886121. https://doi.org/10.1073/pnas.2407886121

Bornmann, L., Haunschild, R., & Mutz, R. (2021). Growth rates of modern science: A latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, *8*(1), 224. https://doi.org/10.1057/s41599-021-00903-w

Bromme, R., & Goldman, S. R. (2014). The Public's Bounded Understanding of Science. *Educational Psychologist*, *49*(2), 59–69. https://doi.org/10.1080/00461520.2014.921572

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS)*, *159*, 1877–1901. https://doi.org/10.5555/3495724.3495883

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–21. https://doi.org/10.1145/3449287

Buder, J., Lindner, M., Oestermeier, U., Huff, M., Gerjets, P., Utz, S., & Cress, U. (2024). Generative Künstliche Intelligenz: Mögliche Auswirkungen auf die psychologische Forschung. *Psychologische Rundschau*. https://doi.org/10.1026/0033-3042/a000699

Burton, J. W., Stein, M.-K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, *33*(2), 220–239. https://doi.org/10.1002/bdm.2155

Carlson, M. (2015). The robotic reporter: Automated journalism and the redefinition of labor, compositional forms, and journalistic authority. *Digit. Journal.*, *3*(3), 416–431. https://doi.org/10.1080/21670811.2014.976412

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, *56*(5), 809–825. https://doi.org/10.1177/0022243719851788

Cave, S., Coughlan, K., & Dihal, K. (2019). "Scary Robots": Examining Public Responses to AI. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 331–337. https://doi.org/10.1145/3306618.3314232

Ciechanowski, L., Przegalinska, A., Magnuski, M., & Gloor, P. (2019). In the Shades of the Uncanny Valley: An Experimental Study of Human-Chatbot Interaction. *Future Generation Computer Systems-the International Journal of Escience*, *92*, 539–548. https://doi.org/10.1016/j.future.2018.01.055

Cloudy, J., Banks, J., & Bowman, N. D. (2022). Ai journalists and reduction of perceived hostile media bias: Replication and extension considering news organization cues. *Technology, Mind, and Behavior*, *3*(3). https://doi.org/10.1037/tmb0000083

Cologna, V., Mede, N. G., Berger, S., Besley, J., Brick, C., Joubert, M., Maibach, E. W., Mihelj, S., Oreskes, N., Schäfer, M. S., Van Der Linden, S., Abdul Aziz, N. I., Abdulsalam, S., Shamsi, N. A., Aczel, B., Adinugroho, I., Alabrese, E., Aldoh, A., Alfano, M., ... Zwaan, R. A. (2025). Trust in scientists and their role in society across 68 countries. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-024-02090-5

Committee on Toward an Open Science Enterprise, Board on Research Data and Information, Policy and Global Affairs, & National Academies of Sciences, Engineering, and Medicine. (2018). *Open Science by Design: Realizing a Vision for 21st Century Research*. National Academies Press. https://doi.org/10.17226/25116

Connolly, J. M., Uscinski, J. E., Klofstad, C. A., & West, J. P. (2019). Communicating to the Public in the Era of Conspiracy Theory. *Public Integrity*, *21*(5), 469–476. https://doi.org/10.1080/10999922.2019.1603045

Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science*, *385*(6714), Article eadq1814. https://doi.org/10.1126/science.adq1814

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *Mis Q.*, *13*(3), 319–340. https://doi.org/10.2307/249008

De Boer, B. (2017). Evolution of speech and evolution of language. *Psychonomic Bulletin & Review*, *24*(1), 158–162. https://doi.org/10.3758/s13423-016-1130-6

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. *Journal of Experimental Psychology-General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Management Science*, *64*(3), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

Eigner, E., & Händler, T. (2024). *Determinants of LLM-assisted Decision-Making*. arXiv. http://doi.org/10.48550/arXiv.2402.17385

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On Seeing Human: A Three-Factor Theory of Anthropomorphism. *Psychological Review*, *114*(4), 864–86. https://doi.org/10.1037/0033-295X.114.4.864

Fletcher, R., & Nielsen, R. K. (2024). *What does the public in six countries think of generative AI in news?* (Tech. rep.). Reuters Institute for the Study of Journalism. https://doi.org/10.60625/RISJ-4ZB8-CG87

Forchtner, B., Kroneder, A., & Wetzel, D. (2018). Being Skeptical? Exploring Far-Right Climate-Change Communication in Germany. *Environmental Communication*, *12*(5), 589–604. https://doi.org/10.1080/17524032.2018.1470546

Ge, X., Xu, C., Misaki, D., Markus, H. R., & Tsai, J. L. (2024). How Culture Shapes What People Want From AI. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3613904.3642660

Géron, A. (2023). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (Third edition). O'Reilly.

Glikson, E., & Woolley, A. W. (2020). Human Trust in Artificial Intelligence: Review of Empirical Research. *Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Goodfellow, I., Courville, A., & Bengio, Y. (2016). *Deep learning*. The MIT Press. https://www.deeplearningbook.org/

Gottschling, M. (2024). Imitationen: Zur Menschlichkeit des Erzählens mit Künstlicher Intelligenz. In A. Burkhardt, S. Marschal, & O. Kramer (Eds.), *Artificial Turn: Interdisziplinäre Perspektiven auf Künstliche Intelligenz* (pp. 1–30). wbg Academic. https://rhet.ai/2023/12/06/imitationen-zur-menschlichkeit-des-erzaehlens-mit-kuenstlicher-intelligenz/

Graefe, A., Journalism, C. U. G. S. o., Journalism, C. U. G. S. o. J. T. C. f. D., & GitBook. (2016). *Guide to Automated Journalism*. Columbia Journalism School. https://books.google.de/books?id=0iPbjwEACAAJ

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

Graefe, A., Haim, M., Haarmann, B., & Brosius, H.-B. (2018). Readers' perception of computer-generated news: Credibility, expertise, and readability. *Journalism*, *19*(5), 595–610. https://doi.org/10.1177/1464884916641269

Greussing, E., Guenther, L., Baram-Tsabari, A., Dabran-Zivan, S., Jonas, E., Klein-Avraham, I., Taddicken, M., Agergaard, T. E., Beets, B., Brossard, D., Chakraborty, A., Fage-Butler, A., Huang, C.-J., Kankaria, S., Lo, Y.-Y., Nielsen, K. H., Riedlinger, M., & Song, H. (2025). The perception and use of generative AI for science-related information search: Insights from a cross-national study. *Public Understanding of Science*, *0*(0). https://doi.org/10.1177/09636625241308493

Grimes, G. M., Schuetzler, R. M., & Giboney, J. S. (2021). Mental models and expectation violations in conversational AI interactions. *Decision Support Systems*, *144*, Article 113515. https://doi.org/10.1016/j.dss.2021.113515

Guzman, A. L., & Lewis, S. C. (2020). Artificial intelligence and communication: A Human-Machine Communication research agenda. *New Media & Society*, *22*(1), 70–86. https://doi.org/10.1177/1461444819858691

Haas, A., & Unkel, J. (2017). Ranking versus reputation: Perception and effects of search result credibility. *Behaviour & Information Technology*, *36*(12), 1285–1298. https://doi.org/10.1080/0144929X.2017.1381166

Haenlein, M., & Kaplan, A. (2019). A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. *California Management Review*, *61*(4), 5–14. https://doi.org/10.1177/0008125619864925

Harambam, J., & Aupers, S. (2015). Contesting epistemic authority: Conspiracy theories on the boundaries of science. *Public Understanding of Science*, *24*(4), 466–480. https://doi.org/10.1177/0963662514559891

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., & Kingsbury, B. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine*, *29*(6), 82–97. https://doi.org/10.1109/MSP.2012.2205597

Hou, Y. T.-Y., & Jung, M. F. (2021). Who is the Expert? Reconciling Algorithm Aversion and Algorithm Appreciation in AI-Supported Decision Making. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–25. https://doi.org/10.1145/3479864

Huang, M.-H., & Rust, R. T. (2018). Artificial Intelligence in Service. *Journal of Service Research*, *21*(2), 155–172. https://doi.org/10.1177/1094670517752459

Hundey, E. J., Olker, J. H., Carreira, C., Daigle, R. M., Elgin, A. K., Finiguerra, M., Gownaris, N. J., Hayes, N., Heffner, L., Roxanna Razavi, N., Shirey, P. D., Tolar, B. B., & Wood-Charlson, E. M. (2016). A Shifting Tide: Recommendations for Incorporating Science Communication into Graduate Training. *Limnology and Oceanography Bulletin*, *25*(4), 109–116. https://doi.org/10.1002/lob.10151

Jolley, D., & Douglas, K. M. (2014). The Effects of Anti-Vaccine Conspiracy Theories on Vaccination Intentions (R. Tripp, Ed.). *PLoS ONE*, *9*(2), Article e89177. https://doi.org/10.1371/journal.pone.0089177

Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why are we Averse Towards Algorithms? A Comprehensive Literature Review on Algorithm Aversion. *Proceedings of the 28th European Conference on Information Systems (ECIS), An Online AIS Conference*. https://aisel.aisnet.org/ecis2020_rp/168?utm_source=aisel.aisnet.org%2Fecis2020_rp%2F168&utm_medium=PDF&utm_campaign=PDFCoverPages

Kahneman, D. (2012). *Thinking, fast and slow*. Penguin.

Kelly, S., Kaye, S.-A., & Oviedo-Trespalacios, O. (2023). What factors contribute to the acceptance of artificial intelligence? A systematic review. *Telematics and Informatics*, *77*, Article 101925. https://doi.org/10.1016/j.tele.2022.101925

Kim, D., & Kim, S. (2018). Newspaper journalists' attitudes towards robot journalism. *Telematics and Informatics*, *35*(2), 340–357. https://doi.org/10.1016/j.tele.2017.12.009

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*, Article 106553. https://doi.org/10.1016/j.chb.2020.106553

Kohring, M., & Matthes, J. (2007). Trust in News Media: Development and Validation of a Multidimensional Scale. *Communication research*, *34*(2), 231–252. https://doi.org/10.1177/0093650206298071

Laakso, M., Welling, P., Bukvova, H., Nyman, L., Björk, B.-C., & Hedlund, T. (2011). The Development of Open Access Journal Publishing from 1993 to 2009 (M. Hermes-Lima, Ed.). *PLoS ONE*, *6*(6), Article e20961. https://doi.org/10.1371/journal.pone.0020961

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. https://doi.org/10.1177/2515245918770963

Lasser, J., Ahne, V., Heiler, G., Klimek, P., Metzler, H., Reisch, T., Sprenger, M., Thurner, S., & Sorger, J. (2020). Complexity, transparency and time pressure: Practical insights into science communication in times of crisis. *Journal of Science Communication*, *19*(05), Article N01. https://doi.org/10.22323/2.19050801

Latar, N. L. (2018). *Robot Journalism: Can Human Journalism Survive?* World Scientific. https://doi.org/10.1142/10913

Lee, C.-J., & Scheufele, D. A. (2006). The Influence of Knowledge and Deference toward Scientific Authority: A Media Effects Model for Public Attitudes toward Nanotechnology. *Journalism & Mass Communication Quarterly*, *83*(4), 819–834. https://doi.org/10.1177/107769900608300406

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, *5*(1). https://doi.org/10.1177/2053951718756684

Lermann Henestrosa, A., Greving, H., & Kimmerle, J. (2023). Automated Journalism: The Effects of AI Authorship and Evaluative Information on the Perception of a Science Journalism Article. *Computers in Human Behavior*, *138*, Article 107445. https://doi.org/10.1016/j.chb.2022.107445

Lermann Henestrosa, A., & Kimmerle, J. (2024a). Understanding and Perception of Automated Text Generation among the Public: Two Surveys with Representative

Samples in Germany. *Behavioral Sciences*, *14*(5), Article 353. https://doi.org/10.3390/bs14050353

Lermann Henestrosa, A., & Kimmerle, J. (2024b). The Effects of Assumed AI vs. Human Authorship on the Perception of a GPT-generated Text. *Journalism and Media*, *5*(3), 1085–1097. https://doi.org/10.3390/journalmedia5030069

Lermann Henestrosa, A., & Kimmerle, J. (2024c). Data Descriptor for "Understanding and Perception of Automated Text Generation among the Public: Two Surveys with Representative Samples in Germany". *Data*, *9*(10), Article 116. https://doi.org/10.3390/data9100116

Lermann Henestrosa, A., & Kimmerle, J. (2025). "Always Check Important Information!" - The Role of Disclaimers in the Perception of AI-generated Content. *Computers in Human Behavior: Artificial Humans*, *4*, Article 100142. https://doi.org/10.1016/j.chbah.2025.100142

Lockey, S., Gillespie, N., Holm, D., & Someh, I. A. (2021). A Review of Trust in Artificial Intelligence: Challenges, Vulnerabilities and Future Directions. https://doi.org/10.24251/HICSS.2021.664

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Longoni, C., & Cian, L. (2022). Artificial Intelligence in Utilitarian vs. Hedonic Contexts: The "Word-of-Machine" Effect. *Journal of Marketing*, *86*(1), 91–108. https://doi.org/10.1177/0022242920957347

Luo, X., Tong, S., Fang, Z., & Qu, Z. (2019). Machines vs. Humans: The Impact of Artificial Intelligence Chatbot Disclosure on Customer Purchases. *Marketing Science*, *38*(6), 937–947. https://doi.org/10.1287/mksc.2019.1192

Ma, D., Akram, H., & Chen, H. (2024). Artificial Intelligence in Higher Education: A Cross-Cultural Examination of Students' Behavioral Intentions and Attitudes. *The International Review of Research in Open and Distributed Learning*, *25*(3), 134–157. https://doi.org/10.19173/irrodl.v25i3.7703

Ma, S., Lei, Y., Wang, X., Zheng, C., Shi, C., Yin, M., & Ma, X. (2023). Who Should I Trust: AI or Myself? Leveraging Human and AI Correctness Likelihood to Promote Appropriate Trust in AI-Assisted Decision-Making. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–19. https://doi.org/10.1145/3544548.3581058

Mahmood, A., Wang, J., Yao, B., Wang, D., & Huang, C.-M. (2025). User Interaction Patterns and Breakdowns in Conversing with LLM-Powered Voice Assistants. *International Journal of Human-Computer Studies*, *195*, Article 103406. https://doi.org/10.1016/j.ijhcs.2024.103406

Mariadassou, S., Klesse, A.-K., & Boegershausen, J. (2024). Averse to what: Consumer aversion to algorithmic labels, but not their outputs? *Current Opinion in Psychology*, *58*, 101839. https://doi.org/10.1016/j.copsyc.2024.101839

McCarthy, J., & Hayes, P. (1981). Some Philosophical Problems from the Standpoint of Artificial Intelligence. In *Readings in Artificial Intelligence* (pp. 431–450). Elsevier. https://doi.org/10.1016/B978-0-934613-03-3.50033-7

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazine*, *27*(4). https://doi.org/10.1609/aimag.v27i4.1904

McCombs, M. E. (2004). *Setting the agenda: The mass media and public opinion*. Polity.

McGowan, A., Gui, Y., Dobbs, M., Shuster, S., Cotter, M., Selloni, A., Goodman, M., Srivastava, A., Cecchi, G. A., & Corcoran, C. M. (2023). ChatGPT and Bard exhibit spontaneous citation fabrication during psychiatry literature search. *Psychiatry Research*, *326*, Article 115334. https://doi.org/10.1016/j.psychres.2023.115334

Mede, N. G., & Schäfer, M. S. (2020). Science-related populism: Conceptualizing populist demands toward science. *Public Understanding of Science*, *29*(5), 473–491. https://doi.org/10.1177/0963662520924259

MeMo:KI. (2022). Dashboard des Meinungsmonitor Künstliche Intelligenz. https://www.cais-research.de/forschung/memoki/

Misselhorn, C. (2019). *Grundfragen der Maschinenethik: Catrin Misselhorn* (4. durchgesehende und überarbeitete Auflage). Reclam Verlag.

Mori, M., MacDorman, K., & Kageki, N. (2012). The Uncanny Valley [From the Field]. *IEEE Robotics & Automation Magazine*, *19*(2), 98–100. https://doi.org/10.1109/MRA.2012.2192811

Nourani, M., King, J., & Ragan, E. (2020). The Role of Domain Expertise in User Trust and the Impact of First Impressions with Intelligent Systems. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *8*, 112–121. https://doi.org/10.1609/hcomp.v8i1.7469

Ooi, K.-B., Tan, G. W.-H., Al-Emran, M., Al-Sharafi, M. A., Capatina, A., Chakraborty, A., Dwivedi, Y. K., Huang, T.-L., Kar, A. K., Lee, V.-H., Loh, X.-M., Micu, A., Mikalef, P., Mogaji, E., Pandey, N., Raman, R., Rana, N. P., Sarker, P., Sharma, A., ... Wong, L.-W. (2025). The Potential of Generative Artificial Intelligence Across Disciplines: Perspectives and Future Directions. *Journal of Computer Information Systems*, *65*(1), 76–107. https://doi.org/10.1080/08874417.2023.2261010

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., ... Zoph, B. (2023). *GPT-4 Technical Report*. arXiv. https://doi.org/10.48550/ARXIV.2303.08774

Park, J., & Woo, S. E. (2022). Who Likes Artificial Intelligence? Personality Predictors of Attitudes toward Artificial Intelligence. *The Journal of Psychology*, *156*(1), 68–94. https://doi.org/10.1080/00223980.2021.2012109

Poliakoff, E., & Webb, T. L. (2007). What Factors Predict Scientists' Intentions to Participate in Public Engagement of Science Activities? *Science Communication*, *29*(2), 242–263. https://doi.org/10.1177/1075547007308009

Poole, D. L., Mackworth, A. K., & Goebel, R. (1998). *Computational intelligence: A logical approach*. Oxford University Press.

Potinteu, A. E., Renftle, D., & Said, N. (2023). *What Predicts AI Usage? Investigating the Main Drivers of AI Use Intention over Different Contexts*. PsyArXiv. https://doi.org/10.31234/osf.io/jvdpe

Rakedzon, T., Segev, E., Chapnik, N., Yosef, R., & Baram-Tsabari, A. (2017). Automatic jargon identifier for scientists engaging with the public and science communication educators (S. Lozano, Ed.). *PLoS ONE*, *12*(8), Article e0181742. https://doi.org/10.1371/journal.pone.0181742

Rosengrün, S. (2021). *Künstliche Intelligenz zur Einführung*. Junius.

Russell, S., & Norvig, P. (2016). *Artificial intelligence: A modern approach* (Third edition). Pearson.

Schäfer, M. S. (2012). Taking stock: A meta-analysis of studies on the media's coverage of science. *Public Understanding of Science*, *21*(6), 650–663. https://doi.org/10.1177/0963662510387559

Schäfer, M. S. (2023). The Notorious GPT: Science communication in the age of artificial intelligence. *Journal of Science Communication*, *22*(02). https://doi.org/10.22323/2.22020402

Schepman, A., & Rodway, P. (2020). Initial validation of the general attitudes towards Artificial Intelligence Scale. *Comput. Hum. Behav. Rep.*, *1*, Article 100014. https://doi.org/10.1016/j.chbr.2020.100014

Schwesig, R., Brich, I., Buder, J., Huff, M., & Said, N. (2023). Using artificial intelligence (AI)? Risk and opportunity perception of AI predict people's willingness to use AI. *Journal of Risk Research*, *26*(10), 1053–1084. https://doi.org/10.1080/13669877.2023.2249927

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424. https://doi.org/10.1017/S0140525X00005756

Searle, J. R. (1990). Is the Brain's Mind a Computer Program? *Scientific American*, *262*(1), 25–31. http://www.jstor.org/stable/24996641

Shanahan, M. (2024). Talking about Large Language Models. *Communications of the ACM*, *67*(2), 68–79. https://doi.org/10.1145/3624724

Shen, Y., Heacock, L., Elias, J., Hentel, K. D., Reig, B., Shih, G., & Moy, L. (2023). ChatGPT and Other Large Language Models Are Double-edged Swords. *Radiology*, *307*(2), Article e230163. https://doi.org/10.1148/radiol.230163

Shulman, H. C., Dixon, G. N., Bullock, O. M., & Colón Amill, D. (2020). The Effects of Jargon on Processing Fluency, Self-Perceptions, and Scientific Engagement. *Journal of Language and Social Psychology*, *39*(5-6), 579–597. https://doi.org/10.1177/0261927X20902177

Sundar, S. S. (2008). The MAIN model: A heuristic approach to understanding technology effects on credibility. In J. M. Metzger & J. A. Flanagin (Eds.), *Digital Media, Youth, and Credibility* (pp. 73–100). The MIT Press. https://betterlegalinfo.ca/wp-content/uploads/2019/12/Sundar-paper.pdf

Sundar, S. S., & Kim, J. (2019). Machine heuristic: When we trust computers more than humans with our personal information. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–9. https://doi.org/10.1145/3290605.3300768

Tandoc, E. C., Yao, L. J., & Wu, S. (2020). Man vs. machine? The impact of algorithm authorship on news credibility. *Digital Journalism*, *8*(4), 548–562. https://doi.org/10.1080/21670811.2020.1762102

Tatalovic, M. (2018). AI writing bots are about to revolutionise science journalism: We must shape how this is done. *Jcom-Journal of Science Communication*, *17*(1), 1–7. https://doi.org/10.22323/2.17010501

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R. (2023). Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW1), 1–38. https://doi.org/10.1145/3579605

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. u., & Polosukhin, I. (2017). Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Venkatesh, Morris, Davis, & Davis. (2003). User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly*, *27*(3), 425–478. https://doi.org/10.2307/30036540

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Manag. Sci*, *46*(2), 186–204. https://doi.org/10.1287/mnsc.46.2.186.11926

Waddell, T. F. (2019). Can an Algorithm Reduce the Perceived Bias of News? Testing the Effect of Machine Attribution on News Readers' Evaluations of Bias, Anthropo-

morphism, and Credibility. *Journalism & Mass Communication Quarterly*, *96*(1), 82–100. https://doi.org/10.1177/1077699018815891

Wang, H., Dang, A., Wu, Z., & Mac, S. (2024). Generative AI in higher education: Seeing ChatGPT through universities' policies, resources, and guidelines. *Computers and Education: Artificial Intelligence*, *7*, Article 100326. https://doi.org/10.1016/j.caeai.2024.100326

Wang, P. (2019). On Defining Artificial Intelligence. *Journal of Artificial General Intelligence*, *10*(2), 1–37. https://doi.org/10.2478/jagi-2019-0002

Wang, S. (2021). Moderating Uncivil User Comments by Humans or Machines? The Effects of Moderation Agent on Perceptions of Bias and Credibility in News Content. *Digital Journalism*, *9*(1), 64–83. https://doi.org/10.1080/21670811.2020.1851279

Wang, S., & Huang, G. (2024). The Impact of Machine Authorship on News Audience Perceptions: A Meta-Analysis of Experimental Studies. *Communication Research*, *51*(7), 815–842. https://doi.org/10.1177/00936502241229794

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. arXiv. https://doi.org/10.48550/ARXIV.2201.11903

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., … Gabriel, I. (2022). Taxonomy of Risks posed by Language Models. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 214–229. https://doi.org/10.1145/3531146.3533088

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45. https://doi.org/10.1145/365153.365168

Yang, H., & Sundar, S. S. (2024). Machine heuristic: Concept explication and development of a measurement scale (A. Joinson, Ed.). *Journal of Computer-Mediated Communication*, *29*(6), Article zmae019. https://doi.org/10.1093/jcmc/zmae019

Zhang, B., & Dafoe, A. (2019). Artificial Intelligence: American Attitudes and Trends. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3312874

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., … Wen, J.-R. (2025). *A Survey of Large Language Models*. arXiv. http://doi.org/10.48550/arXiv.2303.18223

Appendices

A Acknowledgments

Of course, this work represents not only my individual effort, but a collective achievement made possible by the support and dedication of many. Without this incredible network, my dissertation simply would have not been possible.

First of all, special thanks to my supervisor, who placed his trust in me from the very beginning, enabling me to work freely and grow scientifically. Thank you, Joachim, for always helping me find the red thread, not stray too far from the course, and set priorities.

I would also like to thank my second supervisor, Ulrike, who has placed just as much trust in me and has confirmed that what I am doing is relevant and worthwhile.

I am very grateful to have found Hannah as a mentor right at the beginning. In so many conversations, not only about scientific work and methods, but also about self-doubt and motivational holes—she built me up, grounded me, and pushed me to keep going.

Special thanks also go to Maren, Steffi, Nora, Lara, Kevin, and Johannes, whose knowledge and expertise I was able to draw and who always knew the answer to every question—and there are many questions also at the end of a dissertation.

Of course, special thanks are also due to the various anonymous reviewers of my manuscripts for their helpful and constructive feedback, which improved my scientific writing.

Finally, and this is the most important part, because these people were there for me and will continue to have my back no matter where my path takes me, I am infinitely grateful for my family and friends.

Thanks to my friends for shared laughter, for your open ears, for your benign feedback, and for your understanding in hard times. You can't get through any project without friends.

Eternal thanks to my parents, my sister, and my family-in-law, without whose support in all areas this work would be far from finished.

And last but not least, thank you, Tom, for probably believing in me more than I ever will, and thank you, my children, for showing me time and again what really matters. You light up my life.

B Copies of Manuscripts

B.1 Manuscript I

Lermann Henestrosa, A., Greving, H., & Kimmerle, J. (2023). Automated Journalism: The Effects of AI Authorship and Evaluative Information on the Perception of a Science Journalism Article. *Computers in Human Behavior, 138*, Article 107445. https://doi.org/10.1016/j.chb.2022.107445

Automated Journalism: The Effects of AI Authorship and Evaluative Information on the Perception of a Science Journalism Article

Angelica Lermann Henestrosa[a]*, Hannah Greving[a] and Joachim Kimmerle[a,b]

[a]*Knowledge Construction Lab, Leibniz-Institut fuer Wissensmedien, Tübingen, Germany;* [b]*Department of Psychology, Eberhard Karls University, Tübingen, Germany*

*Corresponding author: Angelica Lermann Henestrosa, Leibniz-Institut fuer Wissensmedien, Schleichstraße 6, D-72076 Tübingen, Germany, a.lermann-henestrosa(at)iwm-tuebingen.de

Funding

The research was funded by the Leibniz-Institut fuer Wissensmedien, Tübingen (STB Data Science)

Declarations of interest

None

Automated Journalism: The Effects of AI Authorship and Evaluative Information on the Perception of a Science Journalism Article

Texts produced by artificial intelligence (AI) are becoming increasingly prevalent in digital journalism. Research suggests that these texts do not differ from human-written texts in their perceived credibility or trustworthiness where simple and short text types are concerned. However, it is unclear how AI-written texts beyond simple fact reporting are perceived. Therefore, this research aimed to expand upon the existing literature on automated journalism by investigating the influence of AI authorship (vs. human authorship) and evaluative information presentation (vs. neutral information presentation). The results of three preregistered experimental studies revealed no differences in perceived credibility and trustworthiness between AI-written and human-written texts. However, presenting information in an evaluative way decreased the perception of credibility and trustworthiness. Moreover, the AI was perceived as less anthropomorphic than the human author. The belief in the machine heuristic was stronger for an AI than for a human author, particularly when participants had actually read an article allegedly written by an AI. A pooled analysis across the data of all three studies underpinned the main effect of information presentation. Concluding, we discuss the findings against the background of AI perception theory and suggest implications for future research.

Keywords: automated journalism; artificial intelligence; algorithm; credibility; trustworthiness; information presentation

1 Introduction

Artificial intelligence (AI) developments, especially machine learning, are progressing rapidly and can easily compete with humans in many tasks, including text composition. Current solutions range from simple data extraction to fill in templates to more elaborated approaches for creating narratives (Graefe, 2016). In practice, AI-based algorithms have become increasingly capable of taking over writing tasks, such as producing news for media organizations (e.g., Associated Press, Forbes), summarizing scientific data (e.g., Open Research knowledge Graph), or writing narratives (e.g., GPT-3; see also Dörr, 2016; Graefe, 2016). These types of algorithms originated from *Natural Language Generation* (NLG), a subdomain of computer linguistics. NLG methods make it possible to artificially produce naturally written language that is mostly no longer distinguishable

from texts written by humans (Brown et al., 2020; Köbis & Mossink, 2021). In practice, readers have been exposed to automatically created content for several years (e.g., Associated Press uses Automated Insights' *Wordsmith*). Previously, limited applicational possibilities allowed newspapers to automate only simple reporting like earning reports (e.g., Forbes). More recently, the content offered to the general public has grown, and, in journalism, AI-based applications are increasingly used for news automation. But research on readers' perception of this novel authorship and its potential effects on the evaluation of the communicated message is still in an early stage. Therefore, this study series aimed at further investigating the perception of AI authorship in combination with information presentation with respect to more advanced text types, which AI is increasingly capable of producing.

In the past few years, research on readers' perceptions of automatically produced content and news automation were categorized under the keyword automated journalism (for related terms and a delimitation of them, see Danzon-Chambaud, 2021). This phenomenon can be defined as the generation of journalistic stories through software and algorithms without any human input, except for the initial programming (Carlson, 2015). Regarding the quality of the output, findings suggest that already in the early days of news automation, readers did not seem to be able to distinguish between automatically and human-written texts (Clerwall, 2014). Recent studies that presented participants with actual AI-written texts and human-written content also showed that readers had difficulties detecting the real author (Jung et al., 2017; Wölker & Powell, 2021). Thus, doubts about the language quality of automated content may become more and more negligible due to the constant development in this area. However, aside from the linguistic aspects, it remains mostly unclear how automatically written content and AI-authorship, if transparently declared, is actually perceived and accepted by readers. The research presented here was based on the assumption that the average online news consumer has rarely come into conscious contact with AI-written content and has at most a limited understanding of the underlying mechanisms of automated news generation. Therefore, in our studies, with "AI authorship" we do not refer to a specific software or application but take account of the fact that in practice, if at all, only a short note indicates the AI authorship.

1.1 Perception of AI-written texts

In general, people tend to be *algorithm averse* (Dietvorst et al., 2015). Algorithm aversion is a phenomenon observed during the emergence of algorithms: People preferred to interact with human agents even though algorithms outperformed humans in many tasks (Jussupow et al., 2020). As people's daily experiences have become increasingly digital, this aversion may have decreased. In fact, regarding short and simple texts, algorithms have the advantage of being perceived as writing objectively (Wu, 2020). This notion fits what Sundar and Kim (2019) referred to as the *machine heuristic*, a mental shortcut indicating that an operating machine is perceived as being objective, accurate, and free from ideological bias. For instance, it has been shown that news selected by a machine were rated more favorably compared to news chosen by a journalist (Sundar & Nass, 2001). Generally, such a heuristic can be problematic, leading to reckless reliance on machines for tasks they are not competent for. This overreliance might be reflected in people's credibility and trust ratings but would not necessarily impact their performance (e.g., in terms of understanding or recollection). So far, it is underexplored if and how the machine heuristic takes effect for automated text generation because the extent to which people are aware of these technologies differs greatly. For readers of an automatically generated text, the accuracy of a weather forecast or a stock report might be linked to processes like reliably reporting numbers or analyzing data. Applied to AI-written texts, this suggests that an AI tool as an author may be more acceptable to people when it meets the expectations of the readers by writing about simple fact- and number-based topics (Tandoc et al., 2020). It is largely unclear to which extent this heuristic is cognitively activated by people in varying situations.

Research findings regarding the similarities and differences between simple AI-written and human-written texts focused on the credibility perceptions of the messages, which is undoubtedly a core factor when consuming online news. This addressed the need pointed out by the MAIN Model to examine the technological effects of an algorithm author on credibility perceptions (Sundar, 2008). The findings, however, are somewhat mixed. In particular, when using actual, automatically produced texts, the content was perceived to be more boring, but also more objective (Clerwall, 2014) and credible (Graefe et al., 2018). In contrast, recent studies found no differences in participants' credibility ratings (Jang et al., 2021; Wölker & Powell, 2021) and no differences in their

credibility expectations between the respective authors (Haim & Graefe, 2017). Findings from studies that manipulated the supposed authorship were also rather mixed. They ranged from perceptions of human-written texts as more credible (Graefe et al., 2018; Waddell, 2018) to no differences in trustworthiness and expertise (Van der Kaa & Krahmer, 2014), to perceptions of AI as more credible, objective and balanced (Graefe et al., 2018; Liu & Wei, 2019; Wu, 2020). In a meta-analysis, Graefe and Bohlken (2020) compared 12 experimental and descriptive studies that were conducted between 2017 and 2020. They found advantages for human-written content in terms of quality and readability. Experimental evidence also suggests higher credibility when participants simply were told that a human had written the article. Overall and across various topics, however, the meta-analysis found no differences in the perceived credibility of human and AI-written news. Thus, the findings for simple AI-written texts suggest that people could accept AI as an author for short news reporting. However, the acceptance of AI-written texts that go beyond number- and fact-based news reporting remains largely unexamined.

1.2 Perception of complex AI-written texts

We use the term "complex text" for text types that exceed pure number- and fact-based texts as they are already used in practice, for example, as automated sports reporting. Complex texts can have various characteristics that need to be discussed with respect to automatically produced content. Complex texts may contain a comprehensive contextualization of content. Even though AI might technically outperform humans in the speed and accuracy of written language, genuinely human activities like interpreting, explaining, and evaluating information will not necessarily be perceived as adequate when expressed by AI. Only very little research has been conducted on this characteristic. Tandoc et al. (2020) compared objectively versus evaluatively written news about an earthquake allegedly written by an algorithm, using value-laden words within the evaluative texts. In their study, credibility decreased for the AI when its material was perceived to have been written in an evaluative way. In contrast, there was no difference between the objective and evaluative texts for the human author condition. Similarly, Liu and Wei (2019) varied the level of interpretation in the articles used in their study. They found the algorithm author to be perceived as less knowledgeable but more credible for

the interpretative article version. Thus, the findings on the perception of evaluative content in AI-written texts are sparse and inconsistent.

Further evidence against the unconditional acceptance of human characteristics imitated by AI comes from research where affect or emotion come into play. AI is avoided or perceived as less functional (Huang & Rust, 2018; Luo et al., 2019) when a task involves intuition, affect, or empathy, but this is not the case for objective tasks (Castelo et al., 2019). Longoni and Cian (2020) referred to this phenomenon as the *word-of-machine effect*: AI recommenders are perceived to be more competent for utilitarian realms and functional goals than for hedonistic realms and affective goals. This also fits the view of Castelo et al. (2019) that the perception of AIs is context-dependent. Furthermore, research on the *uncanny valley effect* suggests that people feel uncomfortable when a machine behaves too human-like (e.g., Ciechanowski et al., 2019), due to its lack of any real personal experience, which is seen as fundamental to humans (Gray & Wegner, 2012; Liu & Wei, 2019). These human-like features could also affect the extent to which algorithms are perceived as intelligent or even as human at all. This so-called *anthropomorphism* refers to people's tendency to attribute human characteristics to non-human entities (Epley et al., 2007). Anthropomorphic cues have been investigated concerning uniquely human features like the outer appearance of human beings (de Visser et al., 2016) or voice (Eyssel et al., 2012). So far, there is no research on readers' perceptions of anthropomorphism regarding the imitation of human writing skills. With the constant development in NLG, it becomes more urgent not only to capture if people can differentiate between human- and AI-written text (Köbis & Mossink, 2021), but also to understand which cues in a text make it appear more humanlike in combination with an obvious declaration of AI authorship. The same is true for a specific aspect that was once exclusive to humans, namely intelligence. In the present case, it is about the perceived intelligence of an algorithm compared to the perceived intelligence of a human author. In the age of AI, the question arises as to which factors influence how intelligent an author, be it a human or a machine, is assessed to be and is therefore considered capable of writing about certain content. This is important for the question developed here in that the perceived intelligence of a machine is significantly correlated with animacy (Bartneck, Kanda, et al., 2009).

Complex texts may also be concerned with complicated or even controversial topics. Here, human skills, such as categorizing information on a general level or seeing

findings in a broader context, are often relevant in journalism. This applies to scientific information, for example, which is rarely certain knowledge, but often tentative information that cannot be adequately captured by short reporting (Flemming et al., 2020; Kimmerle et al., 2015). In addition, scientific information has parallels to the text types investigated so far as it is usually also fact- and number-based. However, topics of high interest to the public are often controversial, and communication goes beyond the mere description of facts. For example, in a study by Longoni et al. (2019), participants were reluctant to use healthcare provided by AI due to their assumption that AI was not capable of taking their unique circumstances or characteristics into consideration. Such reluctance may be, on the one hand, understandable when one's own health is at stake. On the other hand, information that is not personally relevant could be prepared through AI in an easily understandable way and made accessible to a broad audience. In an era when information is freely available and often does not follow gatekeeping standards (Sundar, 2008), this could hold great potential for communicating socially relevant topics such as biodiversity or technological innovations.

1.3 The current research

Previous research on automated journalism has only examined texts where algorithms outperform humans in terms of their technical characteristics. This applies to short news reporting about topics where pure numbers and facts are represented. Such information is rarely questioned or tainted with opinion. However, other distinctive aspects become relevant when it comes to communicating information that consists of more complex information like scientific information. Specifically, the audience acceptance of AI could depend on the evaluation of information in the respective setting. Following this line of thought, we assume that previous findings regarding simple text types of perceiving an AI as equivalent or even superior cannot necessarily be applied to other subject areas and text types in which the information is evaluated and contextualized. Therefore, this research aimed to provide a first step toward filling this gap by investigating the perception of AI authorship, in particular of credibility and trustworthiness as key concerns in a digital media environment (Sundar, 2008), on longer text types about content that surpasses pure fact listing in text form. Additionally, by examining how information is presented, we take into account the subject area characteristics of science

reporting and typically human writing style, which is never purely objective reporting of information.

To investigate this assumption, the first study presented here used texts from the subject areas of biodiversity (i.e., the spreading of wolves in Germany) and technology (i.e., the current status of autonomous driving), as both are timely and relevant topics in science communication. For both topics, elaborated texts are needed to capture relevant information. Though these subjects are evidence-based and have a clear factual basis in terms of data and legal specifications, the information can still be contextualized and accompanied by an opinion as people might have different concerns that should be considered. Concerning the topic of wolves, for instance, people's attitudes toward recolonization can be addressed based on evidence from wildlife research while at the same time allowing the evaluation of this evidence in a text. Based on these considerations, we proposed the following hypotheses:

> H1: Participants will rate a text allegedly written by an AI as less credible than a text allegedly written by a human only when the article represents an evaluative presentation of information, but not when the article is written in a neutral way.
>
> H2: Participants will rate a text allegedly written by an AI as less trustworthy than a text allegedly written by a human only when the article represents an evaluative presentation of information, but not when the article is written in a neutral way.
>
> H3: Participants will have stronger intentions to verify the information from a text allegedly written by an AI than from a text allegedly written by a human only when the article represents an evaluative presentation of information, but not when the article is written in a neutral way.

Furthermore, we investigated as an open research question whether there is any effect of AI authorship and evaluative information on participants' acquisition of knowledge from the articles. Table 1 summarizes the study designs of all three studies presented in this paper and the measured variables, respectively.

Table 1

Study designs and measures of Studies 1-3

	Factor	Design	Levels	Measures
Study 1	Authorship	Between	AI	Message credibility
			Human	Trustworthiness
	information presentation	Between	Neutral	User verification strategies
			Evaluative-positive	Anthropomorphism
	text topic	Within	Wolves in Germany	Intelligence
			Autonomous driving	Recognition task
				Attitudes toward content
				Attitudes toward AI
Study 2	Authorship	Between	AI	Message credibility
			Human	Trustworthiness
	information presentation	Between	Neutral	Technical competence
			Evaluative-positive	Anthropomorphism
			Evaluative-negative	Intelligence
				Recognition task
				Attitudes toward content
				Attitudes toward AI
Study 3	Authorship	Between	AI	
			Human	

information presentation	Between	Neutral Evaluative-positive	Message credibility Trustworthiness Machine heuristic Anthropomorphism Intelligence Recognition task Science communication pre + post Attitudes toward content

2 Study 1

2.1 Method

2.1.1 Design and participants

Study 1 was a mixed-design experiment with authorship (human vs. AI) and information presentation (neutral vs. evaluative-positive) as between-group factors and text topic (wolves in Germany & autonomous driving) as a repeated measurement. The presentation of two articles serves as a repeated measurement with the goal of achieving generalizability across specific topics. From the 244 participants who completed the study, 42 participants had to be excluded from the analysis due to predefined exclusion criteria. Exclusion criteria were a failed manipulation check concerning the authorship or a failed attention check concerning the content of the articles. Participation took place via the online recruitment platform *Prolific*, lasted about 45 minutes, and was compensated with 6 Pounds Sterling. The final sample sizes of the three studies, including gender and age distributions, are shown in Table 2.

Table 2

Sample size, gender, and age distributions for Studies 1-3

		Gender			Age	
	N_{total}	Male	Female	Diverse	*M* (*SD*)	Range
Study 1	202	115	85	2	31.55 (10.81)	18-68
Study 2	267	139	121	7	30.19 (9.85)	18-65
Study 3	246	127	115	4	29.00 (7.80)	18-58

2.1.2 Material and procedure

The experiment was accessible only for German speakers who were 18 years old or older. It was introduced as a study about the quality of science journalism articles. Participants first filled in the informed consent form and were asked about their demographics (gender, age). They were then assigned to one of four conditions. In all conditions, participants read two texts, one about wolves in Germany and another about autonomous driving. The order in which participants read the texts was counterbalanced.

To manipulate the factor authorship, the respective author was introduced. Depending on which text was presented first, a male or a female science journalist was

briefly described (name, age, main fields of work) for the *human authorship* condition. The gender of the alleged human authors was varied to avoid inducing gender-specific effects. In the *AI authorship* condition, a computer algorithm (named "AutomatedTXT") was introduced as the author of the text. Additionally, the alleged sources were briefly listed, and participants were told that the articles had been published in a reputable newspaper. The texts used in each authorship condition were, in fact, identical, so that authorship was manipulated by merely labeling the author as either human or AI-based.

To manipulate the factor information presentation, the information within the texts was presented either neutrally or evaluatively-positive. In the *neutral information presentation* condition, simply the facts and statistics about the topic were presented. In the *evaluative-positive information presentation* condition, evaluative words and paraphrases in favor of wolves or autonomous driving (e.g., "exaggerated", "sufficient", or "completely harmless to humans" for the wolf-text and "… would make all our lives … significantly safer" for the autonomous driving text) were added to the information, while the presented data and facts were identical across all conditions. This procedure resulted in four different experimental conditions: 1.) a human/neutral condition, 2.) a human/evaluative-positive condition, 3.) an AI/neutral condition, and 4.) an AI/evaluative-positive condition.

2.1.3 Measures

A manipulation check on authorship and an attention check on the topic followed the article. Additionally, participants rated how evaluative they perceived the article's tone on a single item from 1 (*absolutely neutral*) to 5 (*absolutely evaluative*).

We measured the perceived credibility of the respective text with 19 items (e.g., "fair", "interesting", or "coherent"; $\alpha = .86$) of the *Message Credibility* scale (Appelman & Sundar, 2016; Sundar, 1999).The perceived trustworthiness of the texts was measured with 12 items (e.g., "The information in the text would be verifiable if examined"; $\alpha = .90$) from the *Trust in News Media* scale (Kohring & Matthes, 2004, 2007). Participants judged each item of these scales on a 5-point Likert scale from 1 (*absolutely disagree*) to 5 (*absolutely agree*). To measure the intentions to verify the information given in the text, we used the nine items (e.g., "I would check if the information is up to date"; $\alpha = .87$) of the *User Verification Strategies* instrument (Flanagin & Metzger, 2000), which were measured on 4-point scales ranging from 1 (*very unlikely*) to 4 (*very likely*). We also asked

participants to rate the perceived *anthropomorphism* of the author of the articles with five bipolar items (e.g., "fake – natural"; α = .90) and the perceived *intelligence* of the author with five bipolar items (e.g., "incompetent – competent"; α = .86) on two 5-point scales from the *Godspeed* instrument (Bartneck, Kulić, et al., 2009). Comprehension of the presented information was measured by a recognition task consisting of 14 statements from the text that were either true or false. This procedure was repeated for the second article. Finally, participants could make further comments in an open question field and were debriefed and thanked.

2.1.4 Control variables

2.1.4.1 Attitudes toward algorithms. At the beginning of the experiment, we measured participants' general attitudes toward algorithms in an adapted version of the Computer Attitude Scale (Nickell & Pinto, 1986). All participants had to rate nine statements (e.g., "The use of algorithms is enhancing our standard of living"; α = .82) on a 5-point Likert scale from 1 (*absolutely disagree*) to 5 (*absolutely agree*). The average score was M = 3.82 (SD = 0.57).

2.1.4.2 Attitudes toward the content. Depending on which text was presented first, we either first measured the attitude toward wolves with five items adapted from Treves et al. (2013), e.g., "I think wolves in Germany are good"; α = .80), or the attitude toward autonomous driving (e.g., "Automated driving is a reasonable development"; α = .90). Both attitudes were measured on 5-point Likert scales from 1 (*absolutely disagree*) to 5 (*absolutely agree*). The average score was M = 3.74 (SD = 0.82) for attitude toward wolves and M = 3.75 (SD = 0.86) for attitude toward autonomous driving.

2.2 Results

A Welch *t*-test concerning the item of perceived evaluation of the texts confirmed the information presentation manipulation as successful. There was a significant difference in how evaluative the participants perceived the texts with lower ratings when the information was presented in a neutral way (M = 1.84, SD = 0.66) than when it was

presented in an evaluative way (M = 3.28, SD = 0.87), t(195.72) = 13.30, p < .001, d = 1.87.

We expected interaction effects of the factors authorship and information presentation for perceived credibility (H1), trustworthiness (H2), and verification strategies (H3). We conducted three separate analyses of variance (ANOVA) with the between-subject factors authorship and information presentation for the dependent variables credibility, trustworthiness, and verification strategies. The means and standard deviations of the dependent measures are depicted in Table 3. The corresponding test statistics of the hypothesis tests and for the recognition task can be seen in Table 4. No effects supporting H1, H2, or H3 were found. Instead, main effects of information presentation occurred concerning message credibility and trustworthiness. Independent of the authorship, participants rated the article to be less credible and less trustworthy when it was written in an evaluative-positive way.

Table 3

Means and standard deviations for message credibility, trustworthiness, user verification strategies, and recognition in Studies 1-3 by the factors authorship and information presentation

		Authorship					
		Study 1		Study 2		Study 3	
		AI	Human	AI	Human	AI	Human
		M (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
Dependent variable	Information presentation						
Message credibility	Neutral	3.98 (0.33)	4.04 (0.39)	4.03 (0.41)	4.16 (0.37)	3.92 (0.40)	3.97 (0.46)
	Evaluative-positive	3.89 (0.34)	3.79 (0.35)	3.92 (0.44)	3.94 (0.45)	3.85 (0.56)	3.85 (0.47)
	Evaluative-negative	-	-	3.53 (0.62)	3.48 (0.58)	-	-
Trustworthiness	Neutral	3.87 (0.47)	4.02 (0.39)	3.96 (0.52)	4.05 (0.46)	3.79 (0.53)	3.84 (0.60)
	Evaluative-positive	3.74 (0.43)	3.69 (0.46)	3.86 (0.60)	3.80 (0.63)	3.68 (0.65)	3.65 (0.57)
	Evaluative-negative	-	-	3.36 (0.83)	3.34 (0.74)	-	-
User verification strategies	Neutral	2.45 (0.61)	2.51 (0.60)	-	-	-	-
	Evaluative-positive	2.59 (0.44)	2.63 (0.50)	-	-	-	-
	Evaluative-negative	-	-	-	-	-	-
Recognition	Neutral	20.98 (2.05)	19.89 (2.51)	9.94 (1.39)	10.00 (1.71)	10.09 (1.79)	10.06 (1.60)
	Evaluative-positive	20.19 (2.59)	20.72 (2.08)	9.87 (1.66)	10.05 (1.36)	10.06 (1.58)	10.10 (1.65)
	Evaluative-negative	-	-	9.51 (1.89)	9.51 (1.83)	-	-

Table 4

ANOVA results in Studies 1-3 for the interaction effect between the factors authorship and information presentation and the main effects of authorship and information presentation on message credibility, trustworthiness, user verification strategies, and recognition

Dependent variable			Authorship x Information presentation			Authorship			Information presentation		
		df_{Num}, df_{Den}	*F*	*p*	η_p^2	*F*	*p*	η_p^2	*F*	*p*	η_p^2
Study 1	Message credibility	1, 198	2.55	.112	-	0.17	.684	-	12.51	< .001	.06
	Trustworthiness	1, 198	2.40	.123	-	0.68	.421	-	13.89	< .001	.07
	User verification strategies	1, 198	0.02	.877	-	0.49	.483	-	3.07	.081	-
	Recognition	1, 198	6.10	.014	.03	0.72	.399	-	0.00	.964	-
Study 2	Message credibility	2, 261	0.76	.469	-	0.33	.569	-	35.19	< .001	.21
	Trustworthiness	2, 261	0.39	.680	-	0.00	.950	-	25.09	< .001	.16
	Recognition	2, 261	0.07	.936	-	0.16	.690	-	2.18	.115	-
Study 3	Message credibility	1, 242	0.26	.609	-	0.19	.666	-	2.43	.120	-
	Trustworthiness	1, 242	0.30	.585	-	0.03	.860	-	4.01	.046	.02
	Recognition	1, 242	0.02	.885	-	0.00	.990	-	0.00	.982	-

Note. Adjustments for post-hoc-tests: Bonferroni.

We also analyzed whether the between-subjects factors influenced the recognition of the information given in the texts. Simple comparisons revealed that only in the neutral information presentation conditions the participants recognized more statements correctly when an AI was specified as the author of the texts.

2.2.1 Further analyses

The explorative analysis of perceived anthropomorphism revealed a main effect of the factor authorship, $F(1, 198) = 7.85$, $p = .006$, $\eta_p^2 = .04$. The alleged human author was perceived to be more anthropomorphic ($M = 3.95$, $SD = 0.58$) than the AI ($M = 3.71$, $SD = 0.69$). We also found a main effect for the factor information presentation, $F(1, 198) = 5.65$, $p = .018$, $\eta_p^2 = .01$, with higher anthropomorphism ratings for the authors when the information was presented evaluatively ($M = 3.93$, $SD = 0.58$) than when presented neutrally ($M = 3.72$, $SD = 0.70$).

For perceived intelligence of the authors, the analysis revealed a main effect of the factor information presentation, $F(1, 198) = 5.02$, $p = .026$, $\eta_p^2 = .03$. The authors were perceived to be more intelligent in the neutral information presentation condition ($M = 4.30$, $SD = 0.44$) than in the evaluative condition ($M = 4.16$, $SD = 0.46$). No further effects were found, Fs < 2.18, $p > .141$.

We also exploratively analyzed if the ratings differed between the two texts. Therefore, we conducted a within-subject ANOVA on message credibility with *text* as a within-subject factor, which revealed no main effect of text, $F(1, 198) = 0.87$, $p = .353$. However, both two-way interactions of text and authorship, $F(1, 198) = 4.14$, $p = .043$, $\eta_p^2 = .02$, and text and information presentation, $F(1, 198) = 4.83$, $p = .029$, $\eta_p^2 = .02$, became significant. When the author was specified to be a human, participants perceived the autonomous driving text to be less credible ($M = 3.86$, $SD = 0.47$) than the wolve-text ($M = 3.96$, $SD = 0.45$). Neither a main effect nor interaction effects of the within- and between-subjects factors occurred for trustworthiness, Fs < 1.44, $p > .231$.

Concerning the user verification strategies, we found no interaction effects of the within- and between-subject factors, Fs < 0.12, $p > .545$. However, there was a main

effect of text, with higher ratings for the autonomous driving text (M = 2.66, SD = 0.57) than for the wolve-text (M = 2.57, SD = 0.58), $F(1, 198) = 11.55$, $p < .001$, $\eta_p^2 = .06$.

2.3 Discussion

The results provided no support for the hypotheses. Instead, credibility and trustworthiness decreased independently of authorship when the information was presented in an evaluative way. Unexpectedly, the factor authorship had no effect on the main variables. However, the results from the recognition task revealed significant differences between the alleged authors within the neutral information presentation condition. Participants recognized more statements in the comprehension tasks correctly when the author of the neutral text was specified to be an AI instead of a human. Because of the relatively new experience of reading a long article that was also transparently declared to be written by an AI, participants may have paid more attention to the text or have read it more carefully, which led to higher scores in the recognition tasks.

Further evidence for different perceptions of the authors came from the anthropomorphism results. The alleged human authors were indeed perceived to be more human-like than the AI, which, however, did not influence the credibility or trustworthiness of the texts. In sum, an evaluative-positive presentation of information, even though it corresponded to participants' opinion, led to lower ratings of credibility and trustworthiness of the texts. At the same time, it resulted in higher anthropomorphism but lowered perceived intelligence of the alleged authors.

Although the introduction of an AI as author had no effect on credibility and trustworthiness perceptions, the significantly lower anthropomorphism ratings indicated the impact of the authorship manipulation. Therefore, it does not seem to be irrelevant who wrote a text, as the perception of an AI author to be less human-like could make a difference when it comes to affect-based or moral content, for instance.

A limitation can be seen in merging the variables over both texts. Two different subject areas were used in order to generalize over topics. Even though the prior attitudes were relatively positive for both topics, different levels of understanding, prior knowledge, or differently strong opinions could also have played a role. This is also reflected in the main and interaction effects of text type with authorship and information presentation on message credibility and user verification strategies revealed by the further analysis. Since the articles differed not only in text topic, we cannot rule out unclear

influences due to the use of different subject areas. Moreover, as the positive evaluation of the presented information was in line with participants' relatively positive prior attitudes, it remained unclear which pattern would result if a text had been used that provided negative evaluations of the same information.

3 Study 2

To address the limitations of Study 1, the autonomous driving material was removed in Study 2, and only the text about wolves was used. We also added a third condition to the information presentation factor, that is, an evaluative-negative condition. As participants in Study 1 had a very positive attitude toward wolves and the evaluative-positive condition clearly supported the recolonization of wolves, participants were predominantly confirmed in their attitude. By using an evaluative-negative version of the text, we aimed to investigate if opinion-confirming effects, which have been shown to be present when reading attitudinally congruent information (Taber & Lodge, 2006), could have covered underlying effects. Finally, we removed the user verification strategies scale.

3.1 Methods

3.1.1 Design and participants

The experiment had a 2 (authorship: human vs. AI) x 3 (information presentation: neutral vs. evaluative-positive vs. evaluative-negative) between-factorial design. From the 286 participants who completed the study, 19 participants had to be excluded from the analysis due to the predefined exclusion criteria. The study took place via Prolific with a duration of about 20 minutes and was compensated with 2.6 Pounds Sterling. Participants who had already participated in Study 1 could not participate in Study 2.

3.1.2 Material and procedure

The experimental procedure followed Study 1, but with the addition of a third condition for the information presentation factor. The evaluative-negative information presentation

contained value-laden words against the spreading of wolves, that is, interpretation of the information about the hazard from wolves (e.g., “dangerous”, “unpleasant”, or “critical”).

3.1.3 Control variables

Like in Study 1, we measured participants’ general attitudes toward algorithms at the beginning of the experiment. The average score on the Computer Attitude Scale was M = 3.79 (SD = 0.54).

Likewise, we again measured the participants’ attitudes toward the article’s content. The average score for attitude toward wolves was M = 4.02 (SD = 0.66).

3.2 Results

A one-way ANOVA for the perceived evaluation of the texts showed that the information presentation manipulation was successful. The texts differed significantly from each other in how evaluative the participants perceived them, with lower ratings when the information was presented in a neutral way (M = 1.66, SD = 0.72) than when it was presented in an evaluative-positive (M = 3.11, SD = 1.08), or evaluative-negative way (M = 3.29, SD = 1.07), Welch’s $F(2, 165.96) = 97.22$, $p < .001$, $\eta_p^2 = .37$.

We conducted two ANOVAs to test the predicted interactions of authorship and information presentation for credibility (H1) and trustworthiness (H2). Again, we found a main effect of information presentation on credibility and trustworthiness. Bonferroni adjusted post-hoc-analyses revealed that the text was perceived to be less credible and trustworthy only when it was written in an evaluative-negative way compared to when it was written in a neutral or evaluative-positive way (Table 3).

3.2.1 Further analyses

As in Study 1, we conducted further explorative analyses of the perceived anthropomorphism and intelligence of the authors. There was again a main effect of the factor authorship on anthropomorphism, $F(2, 261) = 10.79$, $p = .001$, $\eta_p^2 = .04$. Participants perceived the alleged human author to be more anthropomorphic (M = 3.92, SD = 0.67) than the alleged AI (M = 3.62, SD = 0.72). Additionally, a main effect of information presentation occurred, $F(2, 261) = 5.82$, $p = .003$, $\eta_p^2 = .04$. Bonferroni adjusted post-hoc-analyses revealed that participants perceived the authors in the

evaluative-positive condition to be more human-like (M = 3.98, SD = 0.70) than in the neutral (M = 3.70, SD = 0.82) and in the evaluative-negative condition (M = 3.62, SD = 0.69), $p < .031$. The neutral and the evaluative-negative conditions did not differ from each other.

For perceived intelligence, we also found main effects of the between factors. The human author was perceived to be more intelligent (M = 4.26, SD = 0.62) than the AI (M = 4.07, SD = 0.74), $F(2, 261) = 5.41$, $p = .021$, $\eta_p^2 = .02$. Furthermore, the authors in the evaluative-negative condition were perceived to be less intelligent (M = 3.82, SD = 0.79) than in the neutral (M = 4.37, SD = 0.59) and in the evaluative-positive condition (M = 4.27, SD = 0.55), $p < .001$. The neutral and evaluative-positive conditions did not differ in perceived intelligence ($p = .899$).

3.3 Discussion

As in Study 1, there were no interaction effects, neither on credibility nor on trustworthiness. Instead, we found lower credibility and trustworthiness ratings only when the information was presented in an evaluative-negative way. As participants' prior attitudes toward wolves were positive, the results can be considered as reflecting a cognitive process that is aimed at confirming one's own opinion. Seen this way, our results are in line with well-researched cognitive biases in information processing (Nickerson, 1998; Stanovich et al., 2013).

Similar to Study 1, the human author was perceived to be more intelligent and more anthropomorphic than the AI. Against the background of the credibility and trustworthiness perceptions, which did not differ between the authors, the differences in intelligence and anthropomorphism ratings reflect participants' ability to discriminate between the human and the AI author. Moreover, both authors were perceived to be more human-like when the information was presented in an evaluative-positive and less intelligent when the information was presented in an evaluative-negative way.

A factor influencing the results in Studies 1 and 2 could be the measurement of attitudes toward algorithms at the beginning of both experiments. Especially the participants in the human author condition, who were not confronted with an AI-based

algorithm at any time, may have been confused by this measurement about the real author of the texts.

4 Study 3

As ratings of credibility and trustworthiness decreased independently of authorship for the evaluative-negative condition and provided no further empirical insights, this condition was not reused in Study 3. Furthermore, the Computer Attitude Scale was removed to prevent potential doubts or confusion about the alleged authorship. In order to make the context of science communication more salient for participants, and to make the subsequent questions about participants' values when consuming scientific information comprehensible, we briefly introduced the term science communication. This enabled us to ensure every participant had at least a basic understanding of science communication. In turn, offering a definition was important for the subsequent assessment of participants' values for the consumption of scientific content. As human-written text can rarely be completely free of some degree of evaluation, opinion, or bias, it is essential to capture people's attitudes toward opinion expression in journalistic articles.

Furthermore, the differences in the perception of anthropomorphism and intelligence between humans and machines require focused investigation. The theorizing outlined above suggests that, if people are presented with an allegedly AI-written text, the machine as an author is perceived to be less anthropomorphic and less intelligent as a source than the human author. Therefore, we expected higher anthropomorphism and intelligence ratings for the human author than for the AI author.

Moreover, following Sundar and Kim's (2019) theoretical conceptualization, we assessed participants' general beliefs in the machine heuristic for both types of authors in the following experiment. As outlined above, the belief in the machine heuristic refers to the assumption that machines, such as AI-based algorithms, would be more reliable than humans when performing a task. Thus, independently from the factors manipulated in the experiment, we expected participants to ascribe these characteristics rather to an

algorithm writing a text than to a human when directly asked for a comparison between both authors. In Study 3, we proposed the following additional hypotheses:

> H4: There will be a main effect of the factor authorship on the perceived anthropomorphism of the author: The alleged journalist will be perceived as more anthropomorphic than the alleged algorithm.
>
> H5: There will be a main effect of the factor authorship on the perceived intelligence of the author: The alleged journalist will be perceived as more intelligent than the alleged algorithm.
>
> H6: The belief in the machine heuristic will be higher for an algorithm as an author than for a journalist as an author.

4.1 Methods

4.1.1 Design and participants

The experiment had a 2 (authorship: human vs. AI) x 2 (information presentation: neutral vs. evaluative-positive) between-factorial design. A total of 276 participants completed the study, from which 30 participants had to be excluded because of a failed manipulation check or attention check. For sample size, gender, and age distribution, see Table 2. The experiment took place via Prolific with a duration of about 25 minutes and was compensated with 3.13 Pounds Sterling. Participants who had participated in Study 1 or 2 could not participate in Study 3.

4.1.2 Procedure and measures

The procedure of Study 3 followed Studies 1 and 2, but with several adjustments. The experiment began with a brief description of the functions and objectives of science communication in general and an emphasis on the role of science journalism within it, which was presented to all participants. Moreover, we assessed participants' values when consuming scientific information a priori and a posteriori on five items. Participants indicated, for example, whether it is important to them "that facts and opinions are clearly separated from each other". The items were measured on 5-point Likert scales ranging

from 1 (*absolutely disagree*) to 5 (*absolutely agree*). Afterward, we presented the attitude measurement, the authorship manipulation, and the wolf text.

Then we asked participants about their belief in the machine heuristic (Sundar & Kim, 2019; Waddell, 2018). For this purpose, we explained in all conditions that science journalism articles could be produced by humans as well as by AI. Subsequently, four items each followed, related to algorithms (α = .73), and related to humans (α = .76) as authors of a text. Participants had to indicate how much they agreed for both types of authors, in general, to be error-free, unbiased, objective, and accurate when writing a text on 5-point Likert scales from 1 (*absolutely disagree*) to 5 (*absolutely agree*). Finally, participants performed the recognition task, were asked for comments, remarks, and criticism on the text in an open question format, and were debriefed and thanked.

4.1.3 Control variable

Following Studies 1 and 2, we measured participants' attitudes toward the content at the beginning of the experiment. The average score for attitude toward wolves was $M = 3.95$ ($SD = 0.67$).

4.2 Results

A Welch *t*-test showed that the information presentation manipulation was successful: There was a difference in the perceived evaluation of the texts with lower ratings when the information was presented in a neutral way ($M = 1.65$, $SD = 0.70$) than when it was presented in an evaluative-positive way ($M = 3.00$, $SD = 1.09$), $t(212.70) = 11.53$, $p < .001$, $d = 1.47$.

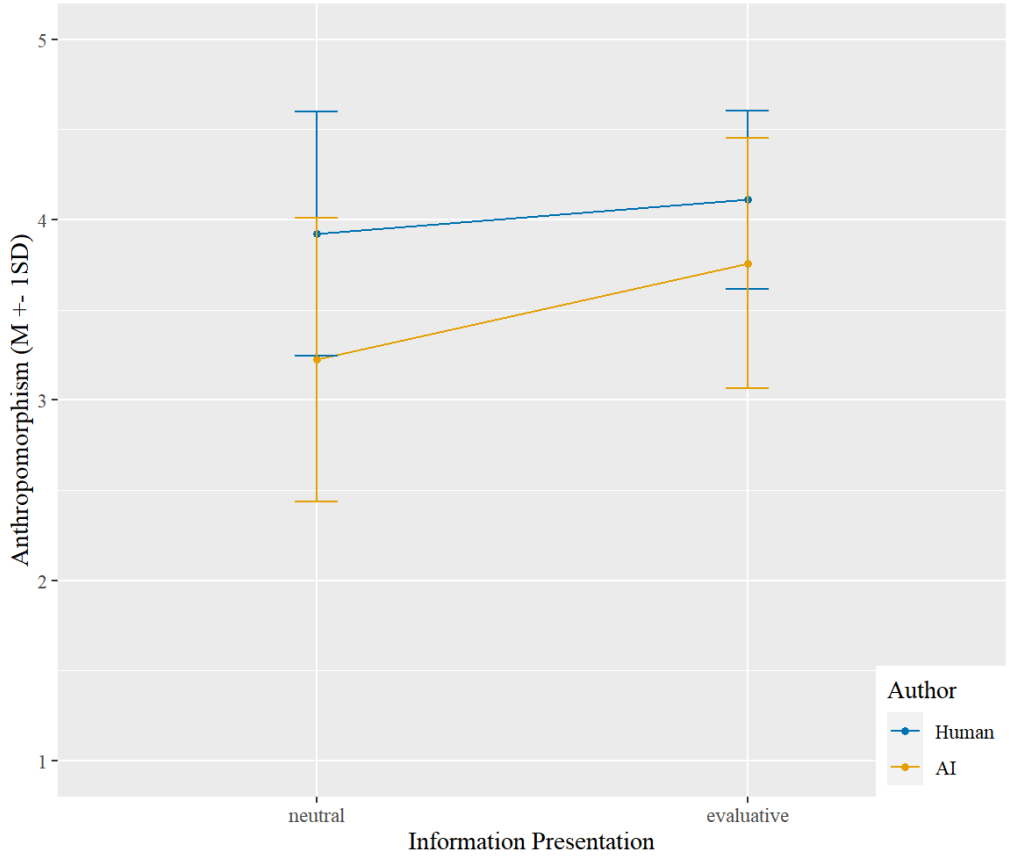
We conducted two ANOVAs to test the predicted interactions of authorship and information presentation for credibility (H1) and trustworthiness (H2). The results of the hypotheses tests concerning message credibility and trustworthiness and for the recognition task are presented in Table 4. For credibility, there were neither an interaction effect nor any main effects. Likewise, no interaction and no main effect for trustworthiness occurred, but we found a main effect for information presentation.

Participants perceived the text to be more trustworthy when it was written in a neutral way than when it was written in an evaluative-positive way.

Concerning perceived anthropomorphism, we hypothesized a main effect of authorship (H4). A Welch *t*-test supported this assumption, and revealed higher ratings for the human author (M = 4.01, SD = 0.60) than for the AI (M = 3.51, SD = 0.78), $t(220.95) = 5.67$, $p < .001$, $d = .72$. To exploratively analyze if authorship and information presentation together had an effect on anthropomorphism, we conducted an ANOVA on this variable. We found a main effect of information presentation on the perceived anthropomorphism, $F(1, 242) = 12.83$, $p < .001$, $\eta_p^2 = .07$. The author was perceived to be more anthropomorphic when the text was written in an evaluative-positive way (M = 3.93, SD = 0.63) than when the text was written neutrally (M = 3.60, SD = 0.81). Moreover, there was an interaction effect of both factors, $F(1, 242) = 4.14$, $p = .043$, $\eta_p^2 = .02$. The AI was perceived to be more anthropomorphic when participants read a text that was written in an evaluative-positive way (M = 3.76, SD = 0.69) than when it was written neutrally (M = 3.23, SD = 0.79), $p < .001$, whereas for the human author, this difference was not significant ($p = .117$). The results are displayed in Figure 1.

Figure 1

Main effects and interaction effect of authorship (human vs. AI) and information presentation (neutral vs. evaluative-positive) on perceived anthropomorphism in Study 3. Error bars represent standard deviations



We also hypothesized a main effect of the factor authorship for perceived intelligence (H5), which was not supported by an unpaired sample *t*-test, $t(244) = 1.85$, $p = .065$. An ANOVA to explore whether authorship and information presentation together had any influence on perceived intelligence revealed no further effects, Fs < 2.07, $p > .152$.

Concerning the machine heuristic, we hypothesized that the belief in the machine heuristic would be stronger for an AI than for a human as the author of a text (H6). A paired sample *t*-test supported this assumption and revealed significantly higher ratings in the machine heuristic items for an AI ($M = 3.42$, $SD = 0.75$) than for a human author ($M = 3.02$, $SD = 0.55$), $t(245) = 7.81$, $p < .001$, $d = .61$. Moreover, we analyzed if the factors authorship and information presentation influenced this pattern. We found an interaction effect with the authorship condition, $F(1, 242) = 9.29$, $p = .003$, $\eta_p^2 = .04$,

indicating that the difference in strength in the belief of the machine heuristic between an AI and a human was even more pronounced when participants had read an article that was allegedly written by an AI than by a person.

4.2.1 Further analyses

The means and standard deviations of participants' answers on their general values when consuming scientific information are shown in Table 5.

Table 5

Means and standard deviations of the five items (a priori) concerning participants' values regarding science communication in Study 3

When it comes to science communication, it is important to me	*M*	*SD*
1) to read objectively written texts.	4.35	0.75
2) to read well-researched texts.	4.81	0.41
3) that the data and facts on a topic are contextualized.	4.33	0.64
4) that the author includes his/her own opinion.	3.57	0.91
5) that facts and opinions are clearly separated.	4.73	0.54

4.3 Discussion

Study 3 aimed to clarify and take the results of Studies 1 and 2 further by emphasizing the context of science communication and measuring participants' values when consuming scientific information. We found no interaction effects of authorship or information presentation concerning credibility and trustworthiness and no main effect of authorship on perceived intelligence. However, our hypothesis that an AI author would be perceived to be less anthropomorphic than a human author was supported. Though the finding that an AI is perceived to be less anthropomorphic is not surprising, it still confirms a clear difference in the perceptions of the alleged authors, which then, however, did not carry through to the assessment of the message.

Moreover, we found evidence for a general belief in the machine heuristic (Sundar & Kim, 2019). Participants explicitly asked for comparing between human and AI-authors believed an AI to be significantly more error-free, unbiased, objective, and accurate when writing a text than a human. Interestingly, this effect was even stronger

when participants had read a complex article that was allegedly written by an AI than when the identical article was allegedly written by a person. Therefore, when the machine heuristic is salient, it could trigger certain expectations toward AI-written texts. Nevertheless, evidence from the data of participants' values when consuming scientific information suggests that readers also expect journalists to be objective. This was more important than the inclusion of the authors' own opinion.

5 Additional Analysis

All of the three studies manipulated *authorship* (human vs. AI) and *information presentation* (neutral vs. evaluative-positive) and used the identical wolf text as a stimulus. Although participants in Study 3 were presented with additional information about science communication in general, and the Computer Attitude Scale was removed in this experiment, the three studies were replications of each other. No changes were made concerning the manipulation of authorship or information presentation. This enabled us to perform two ANCOVAs on the pooled data with study-ID as a covariate for our main variables message credibility and trustworthiness. The results are represented in Table 6.

Table 6

ANCOVA results of the pooled data across Studies 1-3 for the interaction effect between authorship and information presentation and the main effects of authorship and information presentation on message credibility and trustworthiness

Dependent variable	Authorship x Information presentation			Authorship			Information presentation		
	$F(1, 525)$	p	η_p^2	$F(1, 525)$	p	η_p^2	$F(1, 525)$	p	η_p^2
Message credibility	3.23	.073	-	2.97	.085	-	7.10	.008	.01
Trustworthiness	3.62	.058	-	0.68	.421	-	9.58	. 002	.02

Note. The ANCOVA was conducted using only the data of the 2 (authorship: AI vs. human) x 2 (information presentation: neutral vs. evaluative-positive) design with the wolve-text as stimulus in each study (N_{Study1} = 103 for wolve-text at measurement time 1, N_{Study2} = 181, N_{Study3} = 246).

6 General Discussion and Conclusion

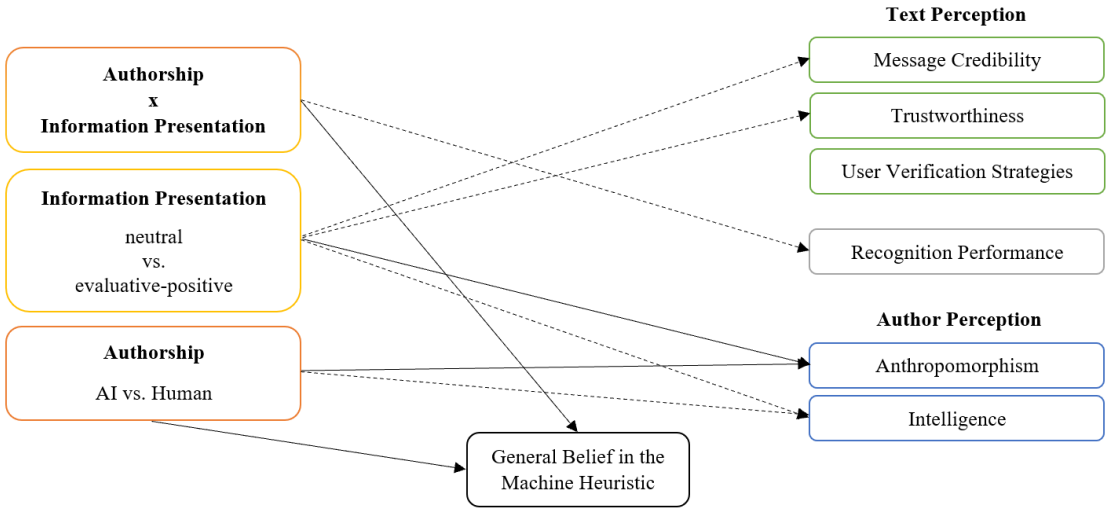
Recently, research on automated journalism has begun to examine readers' perceptions of automatically produced content. The investigated content has been generated using predominantly rule-governed technology targeting a specific output, such as football match reports, even though NLG technologies have become much more powerful. Consequently, various content is already being created automatically, while the effects of AI authorship in general and on more complex text types specifically are under-researched. Therefore, research should consider current developments in automated text generation, such as the improvements made in recent years in large-scale language models (e.g., OpenAI's GPT-3, Microsoft & NVIDIA's Megatron-Turing Natural Language Generation). Such models illustrate the technological capabilities of NLG and the simulation of human written language in general. Although these models are beginning to be used to create specific content in particular application areas, the impact and the perception of automatically generated text and the diversity of possibilities for information presentation on recipients have rarely been considered by researchers so far.

Therefore, the studies presented here aimed at extending the existing literature to text types that exceed pure fact presentation. For more and more topics, structured and machine-readable data is available. Moreover, examples like GPT-3 show that it is possible for AI to write any sort of text. Thus, we conducted three online experiments to investigate the effects of algorithmic authorship and information presentation on the

perception of more complex text types, that is, science journalism articles. Figure 2 depicts the main findings of the three studies taken together.

Figure 2

Graphical overview of the main findings of Studies 1-3 on the measures concerning text perception, recognition performance, author perception, and the general belief in the machine heuristic



Note. Arrows represent significant effects of the factors on the respective variables found across Studies 1-3. Solid arrows represent effects found in all three studies (general belief in the machine heuristic was only measured in Study 3); dashed arrows represent inconsistent findings. For detailed results, see the results sections.

Contrary to our expectations, and even though readers might not be familiar with algorithms writing opinionated texts, credibility and trustworthiness did not decrease when an alleged AI author presented the information about a scientific topic in an evaluative way. Moreover, we did not find any differences between the human and the AI author when the article was written neutrally. Taken together, in all three studies and the additional analysis of the pooled data, which would have been able to detect even very small differences, only information presentation influenced the main dependent variables, regardless of the declared authorship.

This finding is in line with research showing that the use of opinionated language may decrease credibility ratings of messages (Hamilton, 1998; Meyer et al., 2010). In Studies 1 and 3, trustworthiness perceptions were lower when the information was presented in an evaluative-positive way compared to the neutral presentation. In Study 2,

this was only the case when the information was presented in an evaluative-negative way. For credibility, we found a decrease in Studies 1 and 2 in the evaluative conditions but not in Study 3.

Moreover, we found significant differences in the perceived anthropomorphism of the two authors. Participants rated the AI author consistently as less anthropomorphic than the human author. At first glance, this difference might not be surprising and was also found in previous studies (Waddell, 2018, 2019). However, it is interesting when considering the nearly null effects within the credibility and trustworthiness ratings. Although participants made a clear difference in how they perceived the alleged authors, this difference was not at all reflected in their evaluation of the messages written by those authors. Thus, our findings concerning perceived anthropomorphism call into question the previously stated indirect effect of machine attribution on credibility via anthropomorphism (Waddell, 2019). Moreover, they raise the question as to what extent the nature of the author of a text is important to readers for credibility and trustworthiness concerns. Against the background of these findings, the impression emerges that the participants ultimately did not care who wrote a text, even though they were apparently aware that the authors differed from one another. Therefore, the results speak for an acceptance of AI as an author of scientific texts under certain conditions.

It might be that we offered participants an optimal setting by explaining the mode of operation of an allegedly already established AI and its used (reliable) sources. Up to now, there is no requirement to label automatically produced content in journalism, and, in most cases, labels do not go beyond a single byline. However, even when clearly declared and perceived by readers, our findings suggest that it does not make a difference in perceived credibility and trustworthiness for fully informed readers, whether the information comes from a human or an AI.

In addition, these findings are interesting when considering the topic of the articles. Even though participants perceived the AI as less anthropomorphic, they seemed to have no problem with the same AI telling them that wolves were not threatening, for instance, though this obviously concerns human beings more than algorithms. Furthermore, although the credibility and trustworthiness ratings in the evaluative conditions decreased, they did not differ from the human authors. As previous studies have found relationships between anthropomorphism and credibility (Nowak & Rauh, 2005; Waddell, 2018), future studies should further explore this relationship between

perceived anthropomorphism and the credibility of automatically produced journalistic content.

Our studies are among the first to analyze AI authorship on a scientific topic experimentally. So far, research comparing human vs. non-human authorship has focused almost exclusively on news articles about simple data like stocks, sports results, or election polling (e.g., Jung et al., 2017; Waddell, 2018; Zheng et al., 2018), finding small effects at most. Similar to the research presented here, a recent study by Jia and Johnson (2021) using more complex topics like abortion found no significant differences concerning the credibility of stories allegedly written by humans and algorithms. Given that current AI applications for content creation are practically unrestricted in their topics, our approach is a step forward in exploring the perception and evaluation of algorithm authorship. In particular, since human expressions, emotions, and even irony can be simulated and incorporated into artificially generated content, there is a strong need to investigate readers' reactions toward humanoid expressions of AI.

In Studies 1 and 2, we measured participants' attitudes toward algorithms and found that it was relatively positive, indicating a positive basic stance toward developments in automation in general. Additionally, we captured participants' thoughts and remarks at the end of the experiments by open questions. There were only very few participants who doubted the existence of the AI author. At the same time, some comments expressed admiration of the AI instead. However, as a few participants also wondered whether the AI was only performing copy and paste or if a human had preselected the information, future research should further pursue what understanding of AI authorship readers have in mind.

According to Tandoc et al. (2020), a decrease in credibility ratings for an algorithm writing non-objectively aligns with the machine heuristic (Sundar, 2008; Sundar & Kim, 2019). In Study 3, we found evidence for this belief in the machine heuristic, as participants rated an algorithm to be more error-free, unbiased, objective, and accurate than a human author in direct comparison. This effect was even stronger, when participants had indeed read a complex article that they believed was written by an AI compared to an article written by a human being and independently of the information presentation. Even though we found a decrease in credibility and trustworthiness when the AI reported information evaluatively, the same pattern resulted for the human author. As people are used to human writers at least somewhat incorporating their opinions or

evaluations in journalistic articles, the lower credibility and trustworthiness perceptions in the human conditions of the presented studies must be looked at more closely in the future.

These studies have some limitations. The results from the prior attitude scale revealed that the sample was positively disposed toward the chosen topics, and confirmation bias research suggests that messages representing the receivers' views are favored (Festinger, 1957; Nickerson, 1998). As the neutral information presentation condition described facts and current scientific findings also subtly supporting the resettlement of wolves in Germany, participants could have perceived this condition as not inherently neutral, despite its neutral language. Additionally, we could have influenced source credibility perceptions by presenting the participants with sources from which the respective authors had taken their information. We used reliable and generally accepted sources like the Federal Statistical Office to avoid raising any questions about the utilized data, thereby, of course, not fully taking into account the multiplicity of sources embedded in online content (Sundar, 2008). Even though we kept source credibility high to isolate the effect of an AI communicating the information, we cannot rule out that people's implicit assumptions about how humans and algorithms process source information influenced participants' ratings. As little is known about readers' concepts of how a text-producing algorithm works, future studies should consider the difference between the AI as a source, the programmers, and the data the algorithm is fed with, for instance.

All in all, the studies presented here represent a necessary next step in continuing research on automated journalism. By widening the topic and text type scope, our study contributes to and continues previous work by providing a more realistic setting for future applications. Future readers will be more aware of the mode of operation and the use of AI as a potential author, while also questions of labeling requirements might arise. Using a topic with higher complexity and considering the latest developments of NLG, we extended the existing literature on the perception and acceptance of AI-written content toward a more realistic technological future.

7 References

Appelman, A., & Sundar, S. S. (2016). Measuring Message Credibility: Construction and Validation of an Exclusive Scale. *Journalism & Mass Communication Quarterly*, *93*(1), 59-79. https://doi.org/10.1177/1077699015606057

Bartneck, C., Kanda, T., Mubin, O., & Al Mahmud, A. (2009). Does the Design of a Robot Influence Its Animacy and Perceived Intelligence? *International Journal of Social Robotics*, *1*(2), 195-204. https://doi.org/10.1007/s12369-009-0013-7

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics*, *1*(1), 71-81. https://doi.org/10.1007/s12369-008-0001-3

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*.

Carlson, M. (2015). The Robotic Reporter: Automated Journalism and the Redefinition of Labor, Compositional Forms, and Journalistic Authority. *Digital Journalism*, *3*(3), 416-431. https://doi.org/10.1080/21670811.2014.976412

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent Algorithm Aversion. *Journal of Marketing Research*, *56*(5), 809-825. https://doi.org/10.1177/0022243719851788

Ciechanowski, L., Przegalinska, A., Magnuski, M., & Gloor, P. (2019). In the shades of the uncanny valley: An experimental study of human-chatbot interaction. *Future Generation Computer Systems*, *92*, 539-548. https://doi.org/10.1016/j.future.2018.01.055

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

Danzon-Chambaud, S. (2021). A Systematic Review of Automated Journalism Scholarship: Guidelines and Suggestions for Future Research. *Open Research Europe*, *1*(4), 4. https://doi.org/10.12688/openreseurope.13096.1

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A. B., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost Human: Anthropomorphism

Increases Trust Resilience in Cognitive Agents [Article]. *Journal of Experimental Psychology-Applied*, *22*(3), 331-349. https://doi.org/10.1037/xap0000092

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err [Article]. *Journal of Experimental Psychology-General*, *144*(1), 114-126. https://doi.org/10.1037/xge0000033

Dörr, K. N. (2016). Mapping the Field of Algorithmic Journalism [Article]. *Digital Journalism*, *4*(6), 700-722. https://doi.org/10.1080/21670811.2015.1096748

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On Seeing Human: A Three-Factor Theory of Anthropomorphism. *Psychological Review*, *114*(4), 864-886. https://doi.org/10.1037/0033-295X.114.4.864

Eyssel, F., Kuchenbrandt, D., Bobinger, S., de Ruiter, L., Hegel, F., & Assoc Comp, M. (2012, Mar 05-08). 'If You Sound Like Me, You Must Be More Human': On the Interplay of Robot and User Features on Human-Robot Acceptance and Anthropomorphism. *Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction*, Boston, MA.

Festinger, L. (1957). *A Theory of Cognitive Dissonance* (Vol. 2). Stanford university press.

Flanagin, J. A., & Metzger, J. M. (2000). Perceptions of Internet Information Credibility. *Journalism & Mass Communication Quarterly*, *77*(3), 515-540. https://doi.org/10.1177/107769900007700304

Flemming, D., Kimmerle, J., Cress, U., & Sinatra, G. M. (2020). Research is Tentative, but That's Okay: Overcoming Misconceptions about Scientific Tentativeness through Refutation Texts [Article]. *Discourse Processes*, *57*(1), 17-35. https://doi.org/10.1080/0163853x.2019.1629805

Graefe, A. (2016). *Guide to Automated Journalism*. New York, NY: Tow Center for Digital Journalism. https://doi.org/10.7916/D80G3XDJ

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

Graefe, A., Haim, M., Haarmann, B., & Brosius, H.-B. (2018). Readers' Perception of Computer-Generated News: Credibility, Expertise, and Readability. *Journalism*, *19*(5), 595-610. https://doi.org/10.1177/1464884916641269

Gray, K., & Wegner, D. M. (2012). Feeling Robots and Human Zombies: Mind Perception and the Uncanny Valley [Article]. *Cognition*, *125*(1), 125-130. https://doi.org/10.1016/j.cognition.2012.06.007

Haim, M., & Graefe, A. (2017). Automated News: Better than Expected? *Digital Journalism*, *5*(8), 1044-1059. https://doi.org/10.1080/21670811.2017.1345643

Hamilton, M. A. (1998). Message Variables that Mediate and Moderate the Effect of Equivocal Language on Source Credibility. *Journal of Language and Social Psychology*, *17*(1), 109-143. https://doi.org/10.1177/0261927x980171006

Huang, M.-H., & Rust, R. T. (2018). Artificial Intelligence in Service. *Journal of Service Research*, *21*(2), 155-172. https://doi.org/10.1177/1094670517752459

Jang, W., Chun, J. W., Kim, S., & Kang, Y. W. (2021). The Effects of Anthropomorphism on How People Evaluate Algorithm-Written News [Article; Early Access]. *Digital Journalism*, 22. https://doi.org/10.1080/21670811.2021.1976064

Jia, C., & Johnson, T. J. (2021). Source Credibility Matters: Does Automated Journalism Inspire Selective Exposure? *International Journal of Communication*, *15*, 22.

Jung, J., Song, H., Kim, Y., Im, H., & Oh, S. (2017). Intrusion of Software Robots into Journalism: The Public's and Journalists' Perceptions of News Written by Algorithms and Human Journalists. *Computers in human behavior*, *71*, 291-298. https://doi.org/10.1016/j.chb.2017.02.022

Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why are we Averse Towards Algorithms? A Comprehensive Literature Review on Algorithm Aversion. *Proceedings of the 28th European Conference on Information Systems*, Online AIS Conference.

Kimmerle, J., Flemming, D., Feinkohl, I., & Cress, U. (2015). How Laypeople Understand the Tentativeness of Medical Research News in the Media: An Experimental Study on the Perception of Information About Deep Brain Stimulation [Article]. *Science Communication*, *37*(2), 173-189. https://doi.org/10.1177/1075547014556541

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*, 13. https://doi.org/10.1016/j.chb.2020.106553

Kohring, M., & Matthes, J. (2004). Revision und Validierung einer Skala zur Erfassung von Vertrauen in Journalismus. *M&K Medien & Kommunikationswissenschaft*, *52*(3), 377-385. https://doi.org/10.5771/1615-634x-2004-3-377

Kohring, M., & Matthes, J. (2007). Trust in News Media: Development and Validation of a Multidimensional Scale. *Communication research*, *34*(2), 231-252. https://doi.org/10.1177/0093650206298071

Liu, B., & Wei, L. (2019). Machine Authorship in Situ: Effect of News Organization and News Genre on News Credibility [Article]. *Digital Journalism*, *7*(5), 635-657. https://doi.org/10.1080/21670811.2018.1510740

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to Medical Artificial Intelligence [Article]. *Journal of Consumer Research*, *46*(4), 629-650. https://doi.org/10.1093/jcr/ucz013

Longoni, C., & Cian, L. (2020). Artificial Intelligence in Utilitarian vs. Hedonic Contexts: The "Word-of-Machine" Effect. *Journal of Marketing*. https://doi.org/10.1177/0022242920957347

Luo, X., Tong, S., Fang, Z., & Qu, Z. (2019). Frontiers: Machines vs. Humans: The Impact of Artificial Intelligence Chatbot Disclosure on Customer Purchases. *Marketing Science*, *38*(6), 937-947. https://doi.org/10.1287/mksc.2019.1192

Meyer, H. K., Marchionni, D., & Thorson, E. (2010). The Journalist Behind the News: Credibility of Straight, Collaborative, Opinionated, and Blogged "News". *American Behavioral Scientist*, *54*(2), 100-119. https://doi.org/10.1177/0002764210376313

Nickell, G. S., & Pinto, J. N. (1986). The Computer Attitude Scale. *Computers in human behavior*, *2*(4), 301-306. https://doi.org/10.1016/0747-5632(86)90010-5

Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of general psychology*, *2*(2), 175-220. https://doi.org/10.1037/1089-2680.2.2.175

Nowak, K. L., & Rauh, C. (2005). The Influence of the Avatar on Online Perceptions of Anthropomorphism, Androgyny, Credibility, Homophily, and Attraction

[Article]. *Journal of Computer-Mediated Communication*, *11*(1), 26, Article 8. https://doi.org/10.1111/j.1083-6101.2006.tb00308.x

Stanovich, K. E., West, R. F., & Toplak, M. E. (2013). Myside Bias, Rational Thinking, and Intelligence [Article]. *Current Directions in Psychological Science*, *22*(4), 259-264. https://doi.org/10.1177/0963721413480174

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

Sundar, S. S. (2008). The MAIN Model: A Heuristic Approach to Understanding Technology Effects on Credibility. In J. M. Metzger & J. A. Flanagin (Eds.), *Digital Media, Youth, and Credibility* (pp. 73-100). Cambridge, MA: The MIT Press. https://doi.org/1162/dmal.9780262562324.073

Sundar, S. S., & Kim, J. (2019). Machine Heuristic: When We Trust Computers More than Humans with our Personal Information. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow, UK.

Sundar, S. S., & Nass, C. (2001). Conceptualizing Sources in Online News. *Journal of Communication*, *51*(1), 52-72. https://doi.org/10.1093/joc/51.1.52

Taber, C. S., & Lodge, M. (2006). Motivated Skepticism in the Evaluation of Political Beliefs [Article]. *American Journal of Political Science*, *50*(3), 755-769. https://doi.org/10.1111/j.1540-5907.2006.00214.x

Tandoc, E. C., Yao, L. J., & Wu, S. (2020). Man vs. Machine? The Impact of Algorithm Authorship on News Credibility. *Digital Journalism*, *8*(4), 548-562. https://doi.org/10.1080/21670811.2020.1762102

Treves, A., Naughton-Treves, L., & Shelley, V. (2013). Longitudinal Analysis of Attitudes Toward Wolves. *Conservation Biology*, *27*(2), 315-323. https://doi.org/10.1111/cobi.12009

Van der Kaa, H., & Krahmer, E. (2014). Journalist versus News Consumer: The Perceived Credibility of Machine Written News. *Proceedings of the Computation and Journalism Conference*, New York, USA.

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

Waddell, T. F. (2019). Can an Algorithm Reduce the Perceived Bias of News? Testing the Effect of Machine Attribution on News Readers' Evaluations of Bias, Anthropomorphism, and Credibility. *Journalism & Mass Communication Quarterly*, *96*(1), 82-100. https://doi.org/10.1177/1077699018815891

Wölker, A., & Powell, T. E. (2021). Algorithms in the Newsroom? News Readers' Perceived Credibility and Selection of Automated Journalism. *Journalism*, *22*(1), 86-103. https://doi.org/10.1177/1464884918757072

Wu, Y. (2020). 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

Zheng, Y., Zhong, B., & Yang, F. (2018). When Algorithms Meet Journalism: The User Perception to Automated News in a Cross-Cultural Context. *Computers in human behavior*, *86*, 266-275. https://doi.org/10.1016/j.chb.2018.04.046

B.2 Manuscript II

Lermann Henestrosa, A., & Kimmerle, J. (2024). The Effects of Assumed AI vs. Human Authorship on the Perception of a GPT-generated Text. *Journalism and Media, 5*(3), 1085-1097. https://doi.org/10.3390/journalmedia5030069

journalism and media

MDPI

Article

The Effects of Assumed AI vs. Human Authorship on the Perception of a GPT-Generated Text

Angelica Lermann Henestrosa [1,*] and Joachim Kimmerle [1,2]

[1] Knowledge Construction Lab, Leibniz-Institut für Wissensmedien, 72076 Tübingen, Germany
[2] Department of Psychology, Faculty of Science, Eberhard Karls University, 72076 Tübingen, Germany
* Correspondence: a.lermann-henestrosa@iwm-tuebingen.de

Abstract: Artificial Intelligence (AI) has demonstrated its ability to undertake writing tasks, including automated journalism. Prior studies suggest no differences between human and AI authors regarding perceived message credibility. However, research on people's perceptions of AI authorship on complex topics is lacking. In a between-groups experiment (N = 734), we examined the effect of labeled authorship on credibility perceptions of a GPT-written science journalism article. The results of an equivalence test showed that labeling a text as AI-written vs. human-written reduced perceived message credibility (d = 0.36). Moreover, AI authorship decreased perceived source credibility (d = 0.24), anthropomorphism (d = 0.67), and intelligence (d = 0.41). The findings are discussed against the backdrop of a growing availability of AI-generated content and a greater awareness of AI authorship.

Keywords: automated journalism; automated text generation; science communication; credibility; GPT

Citation: Lermann Henestrosa, Angelica, and Joachim Kimmerle. 2024. The Effects of Assumed AI vs. Human Authorship on the Perception of a GPT-Generated Text. *Journalism and Media* 5: 1085–1097. https://doi.org/10.3390/journalmedia5030069

Academic Editors: Rashid Mehmood and João Canavilhas

Received: 4 July 2024
Revised: 6 August 2024
Accepted: 15 August 2024
Published: 20 August 2024

1. Introduction

Automated text generation (ATG) has been garnering significant attention since it became freely available to everyone with internet access due to the release of ChatGPT (Chat-Generative pre-trained transformer). Although ATG has been used for over a decade in areas where structured and machine-readable data was available (e.g., automated journalism), it has long been a niche topic that has received less attention than other developments in artificial intelligence (AI). The availability of large datasets, the increased computational power, advancements in deep learning, and the introduction of the transformer architecture in 2017 (Vaswani et al. 2017), which is the backbone of various models (e.g., GPT from OpenAI or BERT from Google), led to a development boost in natural language generation (NLG) and large language models (LLM). Language, especially written language, is no longer the exclusive preserve of humans. On top of that, due to their training data, LLMs can now write on any topic imaginable. Whether they are also competent in terms of content is a different matter.

Before the emergence of LLMs, ATG had already been an established method in short news reporting, for example. By using structured, machine-readable data, AI-based algorithms can convert the raw data of weather parameters of a city into a consistent verbal weather report, for instance, no more distinguishable from a human-written text (Brown et al. 2020; Köbis and Mossink 2021). Since the release of ChatGPT in November 2022, the possibilities for text generation, even outside journalism, have become apparent. With today's LLMs, tools are at hand that can take away the effort of writing on any topic, only one prompt away. Besides their apparent pitfalls and limitations, they open new possibilities for knowledge access and science communication. For example, scientific information could be made more understandable and approachable by addressing specific target groups, explaining facts to the heart of the matter, summarizing, and breaking down complex information.

Due to the limited capacity of and the scarce attention to ATG before the release of ChatGPT, research is lagging regarding readers' perceptions of this specific form of AI and the perceptions of AI authorship as a novel source cue, especially regarding topics other than news reports. However, studies from the field of automated journalism suggest that at least these texts do not differ from human-written texts in their perceived credibility (Graefe and Bohlken 2020; Jang et al. 2021; Tandoc et al. 2020; Wölker and Powell 2021) or that machine authorship has only a small negative effect on credibility perceptions (Wang and Huang 2024). In a meta-analysis, Graefe and Bohlken (2020) found no difference in credibility perceptions between human and AI authorship across several number- and fact-based topics, such as sports reports or election polling. However, there might be meaningful differences between fully automated news generation, based on structured data, which Graefe and Bohlken took into account, and the support provided by generative AI tools, which has only recently emerged. Besides their finding that assuming an AI vs. a human as an author of a text decreased credibility perceptions in socio-political and environmental topics, Wang and Huang (2024) found no effect on news evaluations in their meta-analysis. However, their findings also revealed a moderating role of the actual source on news evaluation, which could have been due to the quality of the AI texts at the time and might be now less of a problem. On the other hand, Proksch et al. (2024), for example, investigated the effects of labeled authorship on moral topics and found lower author competence and content quality ratings for AI authorship. Also, Böhm et al. (2023) found lower competence ratings for AI-generated content regarding societal and personal challenges simply when the labeled source was AI. These findings indicate a role for the task context and the topic chosen.

While automated journalism research has focused on short news reporting, Lermann Henestrosa et al. (2023) extended this comparison to the topic of science communication. The authors found no differences in message credibility and trustworthiness between humans and AI authors. Still, participants in their study differentiated between the alleged authors regarding perceived anthropomorphism and intelligence, for which the human author was rated significantly higher than the AI author (Lermann Henestrosa et al. 2023). As former research often used only allegedly AI-written texts and was limited to the textual possibilities available at the time, research on actual AI-written content on broader topics is steadily gaining momentum. Moreover, due to small sample sizes in prior studies, small effects could hardly be found and might have led to inconsistent results or null effects.

Therefore, the present study aims to investigate the influence of labeled authorship on the perceived credibility of the message and the credibility of the source of an AI-written science journalistic article. More specifically, our research question was whether there is no difference in the perception of labeled AI authorship vs. human authorship on an actually AI-written scientific article as previous findings suggest (Graefe and Bohlken 2020; Lermann Henestrosa et al. 2023; Wang and Huang 2024). This extends previous research by using an actual AI-written text and expanding the topic to the actual possibilities of LLMs. Using equivalence testing, we responded to the small or non-significant effects of labeled authorship found in prior research. Therefore, we stated the following hypothesis preregistered on https://aspredicted.org/6BP_355 (accessed on 14 August 2024):

H1: *The mean difference in perceived message credibility scores in the two conditions (AI author vs. human author) will be equivalent: the article allegedly written by a human author will be perceived as statistically equally credible as the article allegedly written by an AI.*

Furthermore, previous findings by Lermann Henestrosa et al. (2023) suggest significant differences in authorship perceptions—despite constant material—between alleged AI vs. human authorship, with the AI being perceived as less human-like and less intelligent than the human author. Therefore, we stated the following two hypotheses concerning the perceived anthropomorphism and intelligence of the respective authors:

H2: *There will be a main effect of the factor authorship on the perceived anthropomorphism of the author: the alleged human author will be perceived as more anthropomorphic than the alleged AI.*

H3: *There will be a main effect of the factor authorship on the perceived intelligence of the author: the alleged human author will be perceived as more intelligent than the alleged AI.*

In addition to previous studies, we included the perceived source credibility to directly query the credibility of the alleged author with the following open research question:

RQ1: Is there any effect of labeled authorship on participants' perceived source credibility?

To control for possible effects of attitude toward the text topic, we included the participants' prior attitudes exploratively as covariates.

2. Methods

The study was an online experiment with a one-factorial between-groups design (factor *labeled authorship*: AI author vs. human author). Participants were asked to read a science journalism article about biodiversity, specifically the spread of wolves in Germany, and to rate it on different scales. The exact same text was presented in both conditions, with the only difference being the labeled authorship that was introduced before. The article was created by using the autoregressive language model GPT-3 (OpenAI)—a predecessor model of the model underlying ChatGPT—according to the following procedure: We used the Davinci engine of the OpenAI playground and determined the following parameters: number of tokens = 100, temperature = 0.8, frequency = 0.0, and penalty = 0.0. As prompts, we used five content structuring sentences adapted from the material of Lermann Henestrosa et al. (2023), which were first translated into English and then used to generate five trials per prompt (see Appendix G). From the generated output, a fitting GPT continuation of each of the five trials was selected to gain a complete paragraph, that is, the continuation that most closely matched the tone of a scientific article in terms of content. Thirty-three words had to be deleted because the autocompletion stopped at 100 tokens, resulting in incomplete sentences at the end of the paragraphs (e.g., "In Thuringia" was deleted from the selected autocompletion for the second paragraph). Afterward, the generated text was translated into German and consisted of 353 words. Besides adding three missing punctuation marks, the paragraphs were not edited.

2.1. Sample

A power analysis for a small effect size of d = 0.21, an alpha-error probability of 0.05, and a power of 0.80 revealed a total sample size of N = 714 participants needed for the intended equivalence test for H1. Data from a random and fully anonymized sample were collected via the recruiting platform Prolific. The only prerequisites were that participants had to speak German and be over 18 years old. Of 800 participants in the online experiment, 66 were excluded from the analysis due to preregistered exclusion criteria. The final sample consisted of N = 734 participants with a mean age of 29.04 years old (SD = 9.65). Participation took about 10 min., and participants were compensated with 1.25 GBP. Table 1 shows the absolute and relative distributions for gender and education. For the educational level classification according to school type, see Oehler et al. (2024).

Table 1. Absolute and relative (in percent) numbers of participants by gender and educational level.

	N	%
Gender		
Male	303	41.28
Female	420	57.22
Not specified	11	1.50
Educational level		
Low	4	0.54
Middle	114	15.53
High	616	83.92

2.2. Measures and Procedure

Participants were asked about their demographics (age, gender, educational level) after filling in the informed consent form in the online experiment. As they were told to read a science journalism article about a biodiversity topic, a short introduction to science communication followed, briefly explaining what it is and who practices it. Afterward, their prior *attitude toward wolves* was measured by five items based on a scale (Treves et al. 2013, 5-point Likert scale, e.g., "The spread of wolves in Germany is a positive trend"). Before reading an article, we randomly assigned participants to one of the two conditions, in which the text was either labeled as being written by a human author or by an AI. To make the manipulation clear and prevent participants from thinking that the AI would be taking its information uncontrollably from unclear sources, we briefly explained that the AI could analyze large amounts of data and produce text on this basis without human intervention.

Moreover, as the study was conducted before the launch of ChatGPT, we explained that the AI supposedly took the information for the article by using three reliable sources, which were listed (e.g., the Federal Statistical Office of Germany). The same information was provided for the alleged journalist who was also briefly introduced. The cover story claimed the article was published in 2020 in a German newspaper. For the authorship manipulation, see Appendices A–D. For the original article, see Appendices E and F. After reading the text, participants were asked how neutral they perceived the *tone* of the author to be (bipolar 5-point scale from "absolutely neutral" to "absolutely evaluative") and answered two manipulation check items concerning the author and content of the article.

Dependent variables were the perceived *message credibility* of the text that was measured using the Message Credibility scale (Appelman and Sundar 2016; Sundar 1999) (19 items measured on a 5-point Likert scale). The sample items of this scale are "fair", "accurate", or "authentic". Exploratively, we also measured *source credibility* on five bipolar items, such as "unbiased—biased" or "not trustworthy—trustworthy" (Flanagin and Metzger 2000) (6-point scale). Furthermore, we asked participants to rate the perceived *anthropomorphism* (e.g., "machine-like—human-like") and perceived *intelligence* (e.g., "ignorant—knowledgeable") of the author with five items each (Bartneck et al. 2009) (bipolar 5-point scale). Participants' *attitude toward wolves* was then measured again. Finally, the *behavioral intention* to recommend the article to friends or family and to read such articles again was measured on two single items on a 5-point Likert scale. Before being debriefed, participants could indicate who they thought actually wrote the article by choosing between "I believe the text presented was actually written by an AI", "I believe the text presented was actually written by a human", or "I am not sure" (*true author*).

3. Results

3.1. Main Analyses

The means and standard deviations of all measures by labeled authorship can be seen in Table 2. The explorative analysis of the perceived *tone* of the author revealed no significant difference between the two conditions, Welch-$t(726.05) = 0.30$, $p = 0.762$. Participants perceived the author's tone to be relatively neutral.

Table 2. Means and standard deviations of all measures by labeled authorship.

	Human (n = 359)		AI (n = 375)	
Variable	***M***	***SD***	***M***	***SD***
Author's tone	2.42	0.99	2.41	0.92
Message credibility	3.82	0.53	3.62	0.58
Anthropomorphism	3.80	0.77	3.24	0.88
Intelligence	4.09	0.68	3.80	0.73
Source credibility	4.54	0.96	4.31	1.03
Intention to recommend	3.09	1.16	2.86	1.18
Intention to read	3.75	1.09	3.54	1.12
Prior attitude	3.10	0.38	3.11	0.35
Posterior attitude	3.16	0.36	3.12	0.36

To examine H1, an equivalence test with equivalence bounds of ± the *smallest effect size of interest* (SESOI) of Cohen's d = 0.21 was conducted (1). For a detailed description of equivalence testing, see Lakens et al. (2018). The SESOI was determined based on the raw mean difference of 0.1 on a 5-point scale and the observed pooled variance of $s_p{}^2$ = 0.23 from a previous study.

$$d = \frac{\mu 1 - \mu 2}{\sqrt{s_p{}^2}} = \frac{0.1}{0.48} = 0.21 \quad (1)$$

The equivalence test (TOST [= two one-sided t-tests] procedure) concerning perceived *message credibility* was non-significant, $t(732) = 2.03$, $p = 0.978$, and the observed mean difference of 0.2 fell out of the predefined equivalence bounds. Instead, the null hypothesis significance test (NHST) revealed an effect of labeled authorship on message credibility, $t(732) = 4.87$, $p < 0.001$, $d = 0.36$. The text allegedly written by a human author was perceived as more credible than the same text allegedly written by an AI. The results of the TOST procedure are shown in Figure 1. Figure 2 illustrates the participants' distribution of responses regarding message credibility in each condition.

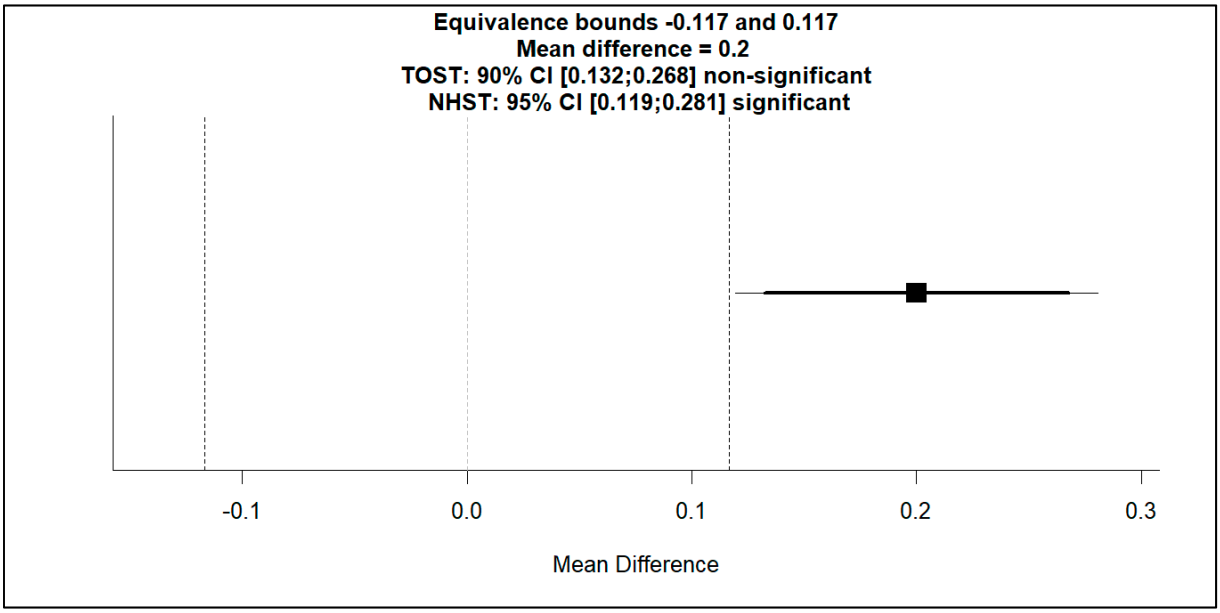


Figure 1. Results of the equivalence test on message credibility. The area between the vertical dashed lines represents the a−priori determined smallest effect size of interest (SESOI).

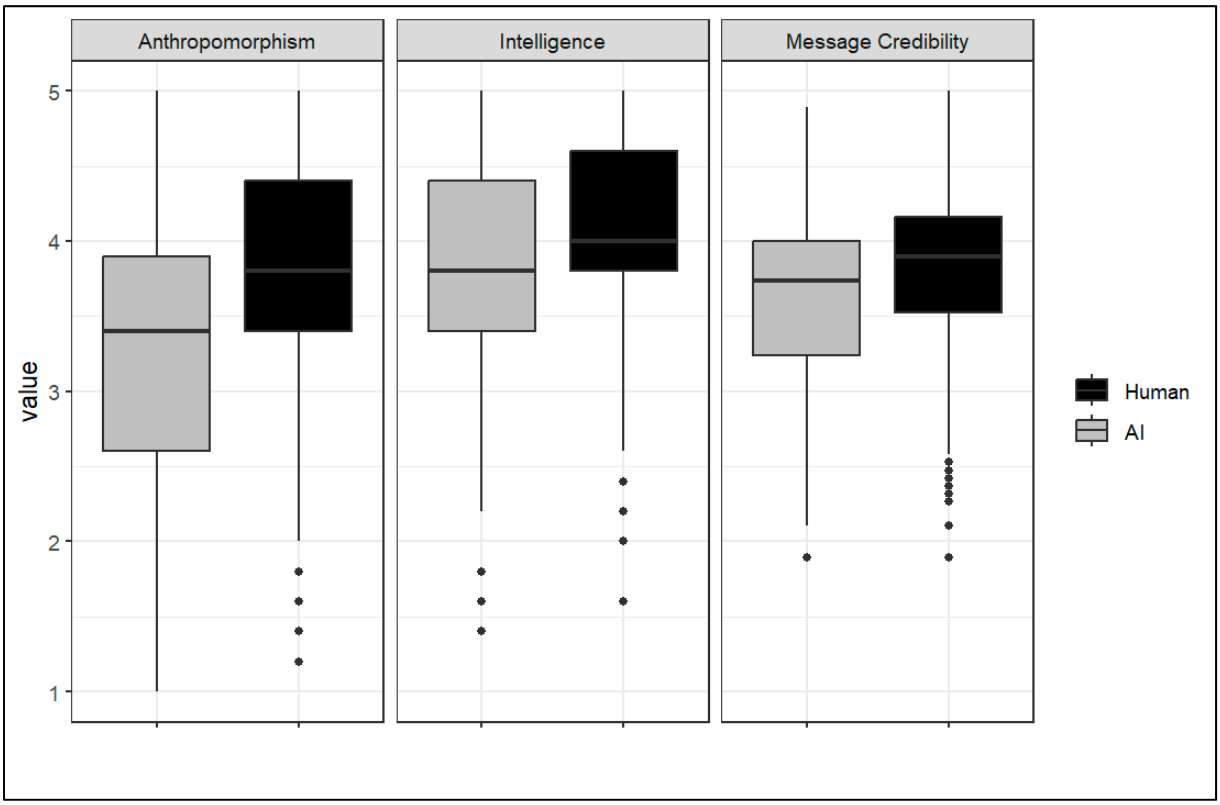


Figure 2. Boxplots of the dependent variables message anthropomorphism, intelligence, and message credibility by labeled authorship.

To examine H2 and H3, two Welch's *t*-tests were conducted due to missing homogeneity of variance. Concerning perceived anthropomorphism, the results revealed a significant difference between the conditions, Welch-$t(726.44) = 9.15$, $p < 0.001$, $d = 0.67$. As expected, and observed in previous studies, participants rated the human author as more anthropomorphic than the AI author. The same pattern resulted regarding perceived intelligence, Welch-$t(731.45) = 5.57$, $p < 0.001$, $d = 0.41$. The human author was perceived as more intelligent than the AI author (see Figure 2).

Regarding the open research question of whether the labeled authorship influenced perceived source credibility, we found a significant difference between the conditions, Welch-$t(731.39) = 3.25$, $p = 0.001$, $d = 0.24$. Participants rated the human author as more credible than the AI author. In addition, source and message credibility were highly correlated, with $r = 0.82$.

Participants also indicated they would recommend the text by the alleged human author more strongly than the one labeled to be written by an AI, Welch-$t(731.41) = 2.70$, $p = 0.007$, $d = 0.20$. Moreover, respondents' intention to read such an article again was higher in the human author condition than in the AI condition, Welch-$t(731.82) = 2.64$, $p = 0.009$, $d = 0.19$. These two items were also highly correlated, with $r = 0.73$

Regarding the *true author* item, 52.65% of the participants in the human author condition believed that a human indeed wrote the text, 21.73% indicated an AI actually wrote it, and 25.63% were unsure. In the AI author condition, 39.47% were convinced that the AI was the actual author, 32.27% thought a human could be the real author, and 28.27% were not sure.

3.2. Further Analyses

Finally, we aimed to explore the prior attitudes toward wolves as a covariate in the analysis. However, we refrained from calculating an ANCOVA because the homogeneity of regression slope was violated concerning message credibility, as the interaction term

was significant, $F(1, 730) = 9.33$, $p < 0.001$, $\eta_p^2 = 0.01$. Figure 3 depicts the interaction effect between participants' prior attitudes toward wolves and labeled authorship, suggesting a positive relationship between participants' initial attitude and perceived message credibility in the human author condition but not in the AI author condition.

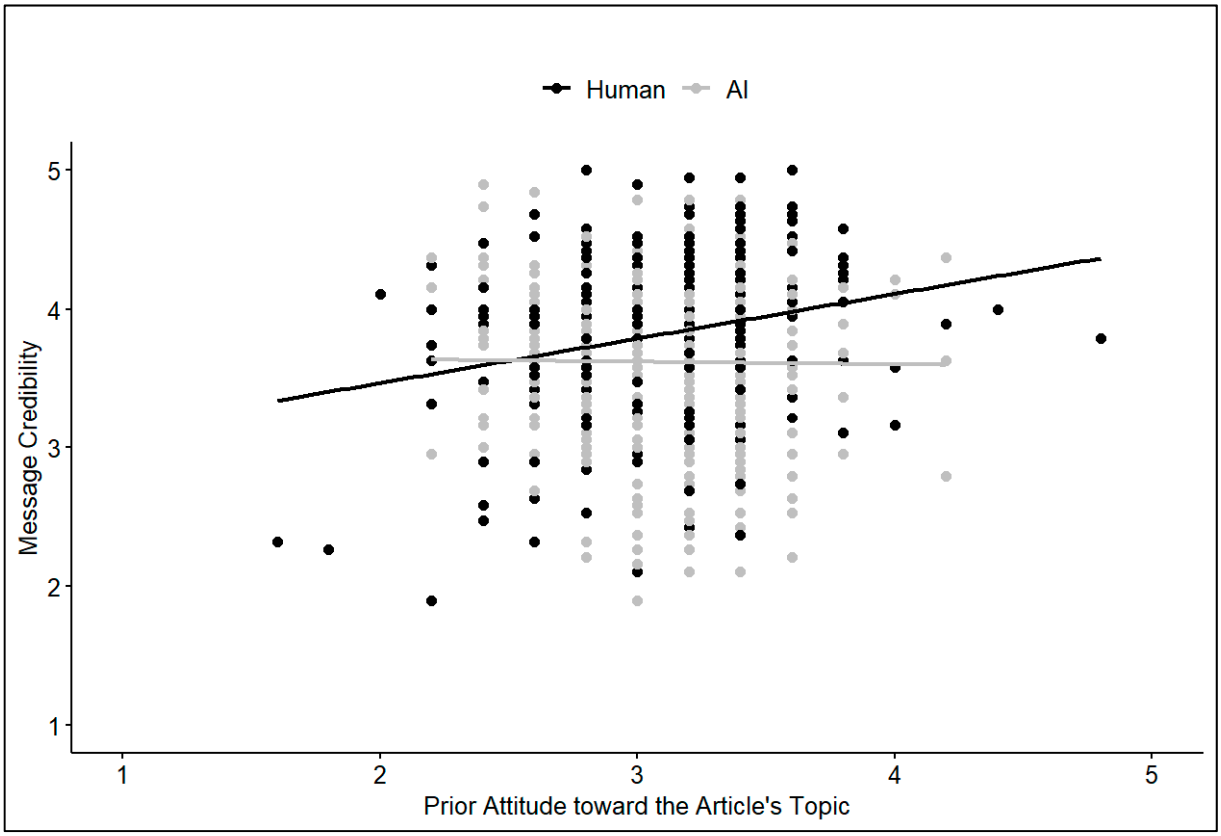


Figure 3. Regression lines depicting the relationship between prior attitude toward the article's topic and perceived message credibility by labeled authorship.

Moreover, an ANOVA with repeated measures regarding participants' attitudes revealed a main effect of treatment, $F(1, 732) = 10.99$, $p < 0.001$, $\eta_p^2 = 0.02$, and a significant interaction effect between treatment and labeled authorship, $F(1, 732) = 4.58$, $p = 0.033$, $\eta_p^2 = 0.01$. Participants' attitudes toward wolves turned out to be more positive after reading the article, especially when a human author had allegedly written the article.

4. Discussion

The study presented here aimed to extend the existing literature on automated journalism to more complex content about science communication. Specifically, we aimed at moving forward research on credibility perceptions of AI authorship on short news reports to a topic and a writing style that better reflects current AI capabilities. Moreover, as one of the first studies in this area, we tested the previously found null effects of human vs. AI authorship on credibility perceptions by using equivalence testing. We defined the smallest effect size of interest and collected a sufficiently large sample size to detect even small effects. Thus, the sufficiently large sample size and the custom-fit methodology to detect even minor differences revealed a small but substantial difference in both perceived message credibility and source credibility between an alleged human and an AI author. The findings show that introducing an AI as the author of a text led to less perceived credibility of the author and the article—even though it was the identical text in both conditions and actually written by an LLM.

In addition, the participants in this experiment were aware of the specific type of author they were dealing with. Besides the passed manipulation check regarding the authorship of all participants included in the analysis, the differences in perceived anthropomorphism and perceived intelligence of the respective authors speak to the differentiating perceptions of the readers: the AI author was rated as less realistic, human-like, intelligent, and competent than the human author. Furthermore, participants intended to recommend the provided article less, and they were less willing to reread such articles in the future when an AI allegedly wrote it. Investigating the factors that led to this evaluation and exploring whether credibility can be manipulated for both types of authors via perceived intelligence and perceived anthropomorphism is a task for future research.

This study is oriented toward current journalistic practice as the text was framed as a science journalism article published in a newspaper. Our experimental setting set a realistic and forward-looking scenario as ATG technology has already been used in journalism for several years and will indeed be deployed much more in the future. Current developments around generative AI, especially around ATG, point to a trend toward the increasing use of AI authorship and its participation in the journalistic process.

Moreover, the provision of scientific information by generative AI reflects actual potential uses of tools such as ChatGPT, also on the part of laypersons. With the increased use of generative AI, its use for information search and information gathering will likely increase. People might trust information provided by ChatGPT in a similar way as that obtained by Google or Wikipedia (Jung et al. 2024). In particular, the more subtle presentation of information in continuous text, when the primary aim may not have been to obtain facts, is still to be investigated and should be viewed critically.

Considering the relatively high credibility ratings for both authors in this experiment, the practical implications of an AI-authored article being perceived as slightly less credible than a human-authored one are unclear. Our study was conducted in January 2022, when peoples' experience with and attitude toward ATG and NLG might not have been highly developed due to missing labeling requirements. In addition, there is evidence that people had no clear concepts regarding ATG and that this has not changed considerably since the release of ChatGPT (Bodani et al. 2023; Lermann Henestrosa and Kimmerle 2024). Given that it was a novel experience for participants to see an AI write about a scientific topic to this extent, the findings are intriguing and speak for a basic leap of faith.

Of course, this study provided very transparent conditions from the readers' perspective by declaring the authorship and the authors' alleged sources, which were reputable. It remains uncertain how transparently media organizations will handle AI in the text-production loop in the future. An extensive societal debate, reflecting the potential desire of readers for clear regulations, is needed to address this concern. For completeness, future studies should also ask for people's assessments of the sources, if provided, and consider today's LLMs' weaknesses in providing reliable sources. In addition, a distinction should be made between LLMs that cannot specify sources due to their basic structure and such models that can do so and would therefore be more appropriate for journalism. However, due to the lack of experience with automatically written journalism, the authorship introduction was necessary for our experiment to successfully manipulate the respective authorship and effectively compare human vs. AI journalism. Therefore, the results of the final question of who the participants ultimately thought was the real author should be interpreted with caution. While in the human condition, they bordered on guessing probability and the proportion of people who believed that an AI had written the text was relatively high given the lack of experience at this time. As technological development is progressing rapidly and debates about the strengths and limitations of ATG are constantly shaping public opinion, this picture might have already changed.

Our exploratory finding was that the credibility of the supposedly human-written article was positively associated with prior attitudes toward wolves, whereas this effect did not extend to the AI author. Additionally, the observation that attitudes toward the topic became more positive after reading the text—particularly with the human author—

suggests that not only could the credibility of human authors be perceived as higher, but their persuasiveness may also be greater. Future studies should further investigate the relationship between attitudes toward content and its perceived credibility, with a focus on how factors such as confirmation bias might influence these dynamics.

With the increased use of ATG technology in journalism, other textual content, and more transparent labeling, future readers will hopefully be much more aware of this novel authorship cue. Moreover, what might complicate or at least change the investigation of AI authorship in future research is the increasing co-authorship and, thus, the blurring of roles in the writing process (Cress and Kimmerle 2023; Luther et al. 2024). A fundamental investigation of the perception of AI authorship is, therefore, long overdue, especially since ChatGPT, but also everybody with the help of ChatGPT and other LLMs, can write about any topic regardless of the truthfulness of the information. Against this background, our results should be seen as a favorable vote of confidence in a technology that holds great potential. On the other hand, they are a warning signal in view of the obvious deficits of LLMs in consistently delivering reliable and robust scientific information.

5. Conclusions

This study contributes to exploring AI authorship on a scientific topic, considering today's LLMs' language and data access capabilities. Prior research suggested small or no differences in message credibility between human and AI authorship, which we tested by using equivalence testing and a sufficiently large sample size to detect even small effects. While the results revealed that the pure suggestion of AI authorship led to lower credibility ratings of the text and author evaluations, major negative effects on credibility were not observed, even on a topic more complex than a weather forecast. Our finding is particularly important in light of the increasing use of LLMs for information searches. Although work is constantly being invested to improve generative AI in terms of factual accuracy, its reliability should be critically scrutinized, especially where scientific information is concerned. More research and a broad public debate are needed to accompany the spread of this specific type of AI and investigate people's attitudes and acceptance of it. In particular, the role of further influencing factors, such as preconceptions and attitudes on generative AI, should be investigated. Since the awareness that an LLM's answers are based on the calculation of probabilities can significantly influence the assessment of the reliability of the information, credibility perceptions might vary with growing experience and attention on this AI area. Therefore, our approach is a step forward in exploring the perception and evaluation of AI authorship for complex topics, reflecting recent developments in NLG.

Author Contributions: Conceptualization, A.L.H. and J.K.; methodology, A.L.H. and J.K.; analysis, A.L.H.; resources, J.K.; data curation, A.L.H.; writing—original draft preparation, A.L.H.; writing—review and editing, A.L.H. and J.K.; visualization, A.L.H.; supervision, J.K.; project administration, J.K.; funding acquisition, J.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Leibniz-Institut für Wissensmedien (STB Data Science).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee of the Leibniz-Institut für Wissensmedien (protocol code LEK 2020/053, approved on 12 November 2020).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The original data presented in the study are openly available in OSF at https://osf.io/gpkc6/ (accessed on 14 August 2024).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Authorship Manipulation—Factor Level "AI" (Original)

Im Folgenden werden Sie einen Text zum Thema Wölfe in Deutschland lesen. Dieser erschien im Frühjahr 2020 auf der Wissenschaftsseite der Süddeutschen Zeitung (SZ).

Er wurde vom Computeralgorithmus *AutomatedTXT* (Version 4.9) verfasst, der Methoden der künstlichen Intelligenz (KI) zur Analyse und Produktion natürlichsprachlicher Texte verwendet. Methoden der künstlichen Intelligenz zur Texterstellung finden bereits seit einigen Jahren Anwendung. *AutomatedTXT* ist so programmiert, dass er eine große Menge an Daten analysieren und die darin enthaltenen Informationen zu einem Text zusammenfügen kann. Eine Überprüfung durch einen Menschen ist dadurch nicht mehr notwendig.

Für den folgenden Text griff *AutomatedTXT* auf öffentlich zugängliche Informationen der Dokumentations- und Beratungsstelle des Bundes für den Wolf (DBBW), des Statistischen Bundesamtes sowie des Bundesministeriums für Umwelt, Naturschutz und nukleare Sicherheit (BMU) zurück.

Appendix B. Authorship Manipulation—Factor Level "AI" (English Translation)

Below, you will read a text on the topic of wolves in Germany. It appeared in spring 2020 on the science page of the Süddeutsche Zeitung (SZ).

It was written by the computer algorithm AutomatedTXT (version 4.9), which uses artificial intelligence (AI) methods to analyze and produce natural language texts. Artificial intelligence methods for text creation have been used for several years. AutomatedTXT is programmed to analyze a large amount of data and combine the information it contains into a text. This means that a human inspection is no longer necessary.

For the following text, AutomatedTXT used publicly available information from the Federal Documentation and Advisory Center for Wolves (DBBW), the Federal Statistical Office, and the Federal Ministry for the Environment, Nature Conservation, and Nuclear Safety (BMU).

Appendix C. Authorship Manipulation—Factor Level "Human" (Original)

Im Folgenden werden Sie einen Artikel zum Thema Wölfe in Deutschland lesen. Dieser erschien im Frühjahr 2020 auf der Wissenschaftsseite der Süddeutschen Zeitung (SZ).

Er wurde von Wissenschaftsjournalist Robert B. Meyer (Jahrgang 1971) verfasst. Seine journalistischen Schwerpunkte liegen in den Bereichen Biodiversität, Naturschutz und Meeresbiologie.

Für den folgenden Text griff der Journalist auf öffentlich zugängliche Informationen der Dokumentations- und Beratungsstelle des Bundes für den Wolf (DBBW), des Statistischen Bundesamtes sowie des Bundesministeriums für Umwelt, Naturschutz und nukleare Sicherheit (BMU) zurück.

Appendix D. Authorship Manipulation—Factor Level "Human" (English Translation)

Below, you will read an article on the topic of wolves in Germany. It appeared in spring 2020 on the science page of the Süddeutsche Zeitung (SZ).

It was written by science journalist Robert B. Meyer (born 1971). His journalistic focus is on the areas of biodiversity, nature conservation, and marine biology.

For the following text, the journalist used publicly available information from the Federal Documentation and Advisory Center for Wolves (DBBW), the Federal Statistical Office, and the Federal Ministry for the Environment, Nature Conservation, and Nuclear Safety (BMU).

Appendix E. Article (Original). Regarding the Authorship Manipulation, an Image of a White, Middle-Aged Man or a Symbol Image for an Algorithm (Program Code) Was Displayed; Not Shown for Copyright Reasons

27 February 2020, 17:08 Uhr
Wölfe in Deutschland

Der Wolf breitet sich in Deutschland wieder aus, eine Art, die vor einem Jahrhundert ein Symbol der Angst war und bis zur Ausrottung gejagt wurde. Die Jagd auf den grauen Wolf ist seit 1945 verboten. Heute leben in Deutschland etwa 150 Wölfe, 75 davon in einem Rudel. Die Art beginnt sich wieder vom Zentrum in die Randgebiete des Landes auszubreiten. In den letzten Jahren wurden Wölfe bis nach Hamburg und München gesichtet, was Naturschützer und Jäger gleichermaßen zur Wachsamkeit aufruft.

Die Zahl der Wölfe in Deutschland wird von der Dokumentations- und Beratungsstelle des Bundes für den Wolf (DBBW) überwacht. In einer am Dienstag veröffentlichten Erklärung teilte die DBBW mit, dass es derzeit zwischen 516 und 680 Wölfe in Deutschland gibt–ein leichter Anstieg gegenüber dem letzten Jahr. Das bedeutet, dass die Population nun den höchsten Stand seit dem 19. Jahrhundert erreicht hat.

Es besteht die Sorge, dass sich der Wolf in Deutschland unkontrolliert ausbreiten und Menschen angreifen könnte. Der erste Wolf in Deutschland kehrte 2012 aus Polen zurück und löste eine Kontroverse aus, nachdem er sechs Schafe im Land getötet hatte. Bayerische Schafhalter betonen, dass sie große Verluste erleiden würden, wenn sich solche Angriffe häufen würden. Wildtierschützer sagen, der Wolf habe ein Recht zu leben und sollte als gefährdete Art geschützt werden.

Die Landwirtschaft fordert Schutzmaßnahmen gegen den Wolf in Deutschland. Zumindest in Teilen Brandenburgs und Sachsens wird der Wolf als Bedrohung für Nutztiere angesehen. Diese Sichtweise erscheint Biologen übertrieben. Der Wolf hat in Deutschland noch nie große oder auch nur mittelgroße Schäden an Nutztieren verursacht, aber er wurde schon bejagt, wenn nur das Gerücht über solche Schäden aufkam.

Es gibt keine Gefahr durch den Wolf in Deutschland. “Wir brauchen keine Wolfsjagd”, sagte ein Sprecher der deutschen Grünen, “die Landwirte können sich selbst schützen.” Greenpeace ist der Meinung, dass die Wolfspopulation als wichtiges Element der Artenvielfalt weiterwachsen sollte, sagte Sprecherin Marie-Christine Keßler gegenüber Reuters. “Wir sind der Meinung, dass der Wolf als Art ein Recht auf Existenz hat”, sagte sie. “Wenn die Behörden ihr Großwild schützen wollen, sollten sie das mit naturverträglichen Mitteln tun und nicht durch das Töten von Tieren.”

Robert B. Meyer/AutomatedTXT

Appendix F. Article (English Translation). Regarding the Authorship Manipulation, an Image of a White, Middle-Aged Man or a Symbol Image for an Algorithm (Program Code) Was Displayed; Not Shown for Copyright Reasons

27 February 2020, 5:08 pm
Wolves in Germany

The wolf is spreading again in Germany, a species that was a symbol of fear a century ago and was hunted to extinction. Hunting the gray wolf has been banned since 1945. Today there are around 150 wolves living in Germany, 75 of them in a pack. The species is beginning to spread again from the center to the outskirts of the country. In recent years, wolves have been spotted as far away as Hamburg and Munich, calling for conservationists and hunters alike to be vigilant.

The number of wolves in Germany is monitored by the Federal Documentation and Advisory Center for Wolves (DBBW). In a statement released on Tuesday, the DBBW said there are currently between 516 and 680 wolves in Germany—a slight increase compared to last year. This means the population is now at its highest level since the 19th century.

There is concern that the wolf could spread uncontrollably in Germany and might attack humans. The first wolf in Germany returned from Poland in 2012 and sparked controversy after killing six sheep in the country. Bavarian sheep farmers emphasize that they would suffer major losses if such attacks became more frequent. Wildlife advocates say the wolf has a right to live and should be protected as an endangered species.

Agriculture calls for protective measures against the wolf in Germany. At least in parts of Brandenburg and Saxony, the wolf is seen as a threat to farm animals. This view seems exaggerated to biologists. The wolf has never caused large or even medium-sized damage to livestock in Germany, but it has been hunted whenever rumors of such damage arose.

There is no danger posed by the wolf in Germany. "We don't need wolf hunting," said a spokesman for the German Green Party, "farmers can protect themselves". Greenpeace believes the wolf population should continue to grow as an important element of biodiversity, spokeswoman Marie-Christine Keßler told Reuters. "We believe that the wolf as a species has a right to exist," she said. "If the authorities want to protect their big game, they should do so using nature-friendly means and not by killing animals."

Robert B. Meyer/AutomatedTXT

Appendix G. Prompts Entered in GPT-3 Playground

1.	The wolf is spreading again in Germany
2.	The number of wolves in Germany is monitored by the Federal Documentation and Advisory Service for the Wolf (DBBW).
3.	There is concern that the wolf could spread uncontrollably in Germany and might attack humans.
4.	Agriculture calls for protection measures against the wolf in Germany.
5.	There is no danger posed by the wolf in Germany

References

Appelman, Alyssa, and Shyam S. Sundar. 2016. Measuring Message Credibility: Construction and Validation of an Exclusive Scale. *Journalism & Mass Communication Quarterly* 93: 59–79. [CrossRef]

Bartneck, Christoph, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics* 1: 71–81. [CrossRef]

Bodani, Nikita, Abhishek Lal, Afsheen Maqsood, Sara Altamash, Naseer Ahmed, and Artak Heboyan. 2023. Knowledge, Attitude, and Practices of General Population Toward Utilizing ChatGPT: A Cross-Sectional Study. *SAGE Open* 13: 21582440231211079. [CrossRef]

Böhm, Robert, Moritz Jörling, Leonhard Reiter, and Christoph Fuchs. 2023. People Devalue Generative AI's Competence but Not Its Advice in Addressing Societal and Personal Challenges. *Communications Psychology* 1: 32. [CrossRef]

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda Askell. 2020. Language Models Are Few-Shot Learners. *arXiv* arXiv:2005.14165.

Cress, Ulrike, and Joachim Kimmerle. 2023. Co-Constructing Knowledge with Generative AI Tools: Reflections from a CSCL Perspective. *International Journal of Computer-Supported Collaborative Learning* 18: 607–14. [CrossRef]

Flanagin, J. Andrew, and J. Miriam Metzger. 2000. Perceptions of Internet Information Credibility. *Journalism & Mass Communication Quarterly* 77: 515–40. [CrossRef]

Graefe, Andreas, and Nina Bohlken. 2020. Automated Journalism: A Meta-Analysis of Readers' Perceptions of Human-Written in Comparison to Automated News. *Media and Communication* 8: 50–59. [CrossRef]

Jang, Wonseok, Jung W. Chun, Soojin Kim, and Young W. Kang. 2021. The Effects of Anthropomorphism on How People Evaluate Algorithm-Written News. *Digital Journalism* 22: 103–24. [CrossRef]
Jung, Yongnam, Cheng Chen, Eunchae Jang, and S. Shyam Sundar. 2024. Do We Trust ChatGPT as Much as Google Search and Wikipedia? In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. Honolulu: ACM, pp. 1–9. [CrossRef]
Köbis, Nils, and Luca D. Mossink. 2021. Artificial Intelligence versus Maya Angelou: Experimental Evidence That People Cannot Differentiate AI-Generated from Human-Written Poetry. *Computers in Human Behavior* 114: 13. [CrossRef]
Lakens, Daniël, Anne M. Scheel, and Peder M. Isager. 2018. Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science* 1: 259–69. [CrossRef]
Lermann Henestrosa, Angelica, and Joachim Kimmerle. 2024. Understanding and Perception of Automated Text Generation among the Public: Two Surveys with Representative Samples in Germany. *Behavioral Sciences* 14: 353. [CrossRef]
Lermann Henestrosa, Angelica, Hannah Greving, and Joachim Kimmerle. 2023. Automated Journalism: The Effects of AI Authorship and Evaluative Information on the Perception of a Science Journalism Article. *Computers in Human Behavior* 138: 107445. [CrossRef]
Luther, Teresa, Joachim Kimmerle, and Ulrike Cress. 2024. Teaming up with an AI: Exploring Human–AI Collaboration in a Writing Scenario with ChatGPT. *AI* 5: 1357–76. [CrossRef]
Oehler, Felicitas, Sophia Kimmig, Robert Hagen, Joachim Kimmerle, Ulrike Cress, Klaus Hackländer, Janosch Arnold, Danny Flemming, and Miriam Brandt. 2024. The Role of Information Presentation for Wildlife Knowledge, Attitude, and Risk Perception. *Conservation Science and Practice* 6: e13089. [CrossRef]
Proksch, Sebastian, Julia Schühle, Elisabeth Streeb, Finn Weymann, Teresa Luther, and Joachim Kimmerle. 2024. The Impact of Text Topic and Assumed Human vs. AI Authorship on Competence and Quality Assessment. *Frontiers in Artificial Intelligence* 7: 1412710. [CrossRef]
Sundar, Shyam S. 1999. Exploring Receivers' Criteria for Perception of Print and Online News. *Journalism & Mass Communication Quarterly* 76: 373–86. [CrossRef]
Tandoc, Edson C., Lim J. Yao, and Shangyuan Wu. 2020. Man vs. Machine? The Impact of Algorithm Authorship on News Credibility. *Digital Journalism* 8: 548–62. [CrossRef]
Treves, Adrian, Lisa Naughton-Treves, and Victoria Shelley. 2013. Longitudinal Analysis of Attitudes Toward Wolves. *Conservation Biology* 27: 315–23. [CrossRef]
Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *Advances in Neural Information Processing Systems* 30. Available online: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (accessed on 14 August 2024).
Wang, Sai, and Guanxiong Huang. 2024. The Impact of Machine Authorship on News Audience Perceptions: A Meta-Analysis of Experimental Studies. *Communication Research*, 00936502241229794. [CrossRef]
Wölker, Anja, and Thomas E. Powell. 2021. Algorithms in the Newsroom? News Readers' Perceived Credibility and Selection of Automated Journalism. *Journalism* 22: 86–103. [CrossRef]

B.3 Manuscript III

Lermann Henestrosa, A., & Kimmerle, J. (2024). Understanding and Perception of Automated Text Generation among the Public: Two Surveys with Representative Samples in Germany. *Behavioral Sciences, 14*(5), Article 353. https://doi.org/10.3390/bs14050353

behavioral sciences

MDPI

Article

Understanding and Perception of Automated Text Generation among the Public: Two Surveys with Representative Samples in Germany

Angelica Lermann Henestrosa [1,*] and Joachim Kimmerle [1,2]

[1] Knowledge Construction Lab, Leibniz-Institut für Wissensmedien, 72076 Tübingen, Germany; j.kimmerle@iwm-tuebingen.de
[2] Department of Psychology, Eberhard Karls University, 72076 Tübingen, Germany
* Correspondence: a.lermann-henestrosa@iwm-tuebingen.de

Abstract: Automated text generation (ATG) technology has evolved rapidly in the last several years, enabling the spread of content produced by artificial intelligence (AI). In addition, with the release of ChatGPT, virtually everyone can now create naturally sounding text on any topic. To optimize future use and understand how humans interact with these technologies, it is essential to capture people's attitudes and beliefs. However, research on ATG perception is lacking. Based on two representative surveys (March 2022: n_1 = 1028; July 2023: n_2 = 1013), we aimed to examine the German population's concepts of and attitudes toward AI authorship. The results revealed a preference for human authorship across a wide range of topics and a lack of knowledge concerning the function, data sources, and responsibilities of ATG. Using multiple regression analysis with k-fold cross-validation, we identified people's attitude toward using ATG, performance expectancy, general attitudes toward AI, and lay attitude toward ChatGPT and ATG as significant predictors of the intention to read AI-written texts in the future. Despite the release of ChatGPT, we observed stability across most variables and minor differences between the two survey points regarding concepts about ATG. We discuss the findings against the backdrop of the ever-increasing availability of automated content and the need for an intensive societal debate about its chances and limitations.

Keywords: automated text generation; public attitudes toward AI; ChatGPT impact; automated journalism

Citation: Lermann Henestrosa, A.; Kimmerle, J. Understanding and Perception of Automated Text Generation among the Public: Two Surveys with Representative Samples in Germany. *Behav. Sci.* **2024**, *14*, 353. https://doi.org/10.3390/bs14050353

Academic Editors: Gengfeng Niu and Xiaochun Xie

Received: 21 February 2024
Revised: 19 April 2024
Accepted: 22 April 2024
Published: 23 April 2024

1. Introduction

Public awareness of the terms automated text generation (ATG), natural language generation (NLG), or large language models (LLM) has grown. The November 2022 release of ChatGPT—a language model developed by the company OpenAI with the capability of generating human-like text in a conversational manner—has fueled the attention to this subfield of artificial intelligence (AI). In detail, NLG is "the subfield of artificial intelligence and computational linguistics that is concerned with the construction of computer systems that can produce understandable texts [. . .] from some underlying non-linguistic representation of information" [1,2]. While the output is always text, the input can vary substantially [3], including flat semantic representations, numerical data, or structured knowledge bases [4]. ATG is a particular type of NLG where natural-sounding text is generated through algorithmic processes with limited human intervention [5]. Already before ChatGPT's release, these automatically produced texts were no longer distinguishable on a linguistic level from those written by humans [6,7]. What is new, however, is that through the wider availability of ATG technology for a broad population, written text will be more and more automatically created. Moreover, unlike previous customer service chatbots, users interact with open access LLMs such as ChatGPT without a particular purpose. Instead, it is a communicative object in itself [2], providing language translation, summarization, or question answering,

all in one tool [8]. Additionally, the underlying technology makes ChatGPT's output unique compared to automatically written text which existed before. Based on a vast amount of data, LLMs have learned how people use written language and how writing works on a statistical level. This makes applications like ChatGPT specifically different from previous ones, such as those used in automated journalism. Due to the fact that the training data are not stored and the output generation does not follow transparently predetermined rules, it is neither predictable nor fully controllable. Consequently, what defines LLMs and leads to the high quality of their output is also one of their most significant weaknesses and dangers for potential users unaware of the genesis of the resulting text.

Among the first to apply ATG technology systematically were news media organizations that used the more rule-based approaches of ATG to automate news reporting (e.g., Associated Press, Forbes, Washington Post). However, the e-commerce sector was also fast to recognize the potential for automating product descriptions, for instance. This shows that AI-generated content has already existed on the internet for over a decade, but the public perception and awareness of this AI subfield has hardly been investigated. However, the release of ChatGPT threw a spotlight on the technological possibilities of generating human-sounding language. GPT-4 (the underlying model of ChatGPT) and other language models like BERT (Google) or XLNet (Microsoft) now have the potential to revolutionize the way people write, perceive, and use text in all contexts. A prerequisite for realizing the full potential of these AI technologies is that users and readers accept and adopt them [9]. While speech-based applications like voice assistants have been known to many for several years, more creative approaches of NLG on more complex topics and data are not very common and hardly salient in public (e.g., Open Research Knowledge Graph). Currently, there is a discrepancy between small groups of people firmly dealing with the benefits of LLMs in general (ranging from individuals revising their complete working process to companies implementing NLG wherever possible) and a large population for whom this technology still is not a reality. However, unlike other AI subfields, such as in the medical context, ATG is no longer just a niche topic for researchers or developers. No more does it concern only those actively using the technology. The amount of AI-generated content on the internet is growing, with some forecasting that the quantity of synthetically generated content will be up to 90% by 2026 [10]. As AI content will become more prevalent in print media, too, consumers will increasingly encounter automatically written text, often without realizing it, especially online. Therefore, it is crucial and unavoidable that consumers deal with AI authorship and develop pertinent opinions. However, a fundamental issue is the general population's lack of awareness of automated texts due to inadequate labeling requirements: it is mostly not labeled or only identified by a single byline. To date, this makes it questionable if and to what extent readers have perceived content to be automatically generated.

In the last decade, many (primarily qualitative) investigations have dealt with journalists' perspectives on this technology, often through job replacement scenarios [11–14]. Investigations into readers' perceptions, attitudes, and beliefs concerning automatically produced content are rare. Consequently, a thoughtful debate about societal, ethical, economic, and juridical implications is late in coming [15]. In short, with the release of ChatGPT, a highly developed technology is accessible to virtually everyone with internet access. At the same time, potential users have not had the time to foster a sharpened awareness of the chances and risks this technology brings.

1.1. Automated Journalism

Before the release of ChatGPT, one of the most popular ATG application areas was the automation of news reporting, also known as automated journalism or robot journalism. Several studies have investigated readers' perceptions and acceptance of automated short news and the novel source cue "AI authorship". However, the findings concerning the perceived credibility of the content and the author are inconsistent. Some studies found that readers perceived AI-written texts as less credible [16], readable [17], and accurate [18].

Others found AI vs. human written texts to be perceived as equal in expertise, trustworthiness [19], and credibility [20–24]. Moreover, some studies found that AI-written texts were perceived as more credible, objective, and balanced [6,17,25] than human-written texts. However, a meta-analysis across different topics of short news reporting shows that differences, if found, were relatively small [26].

These studies have mainly investigated very number-based topics like weather forecasts, earthquake information, or financial reports, for instance. However, the topics' content and reporting style might be essential to the findings. The information presented in a weather forecast or a product description leaves little room for interpretation and stylistic creativity. Advantages of algorithms like accuracy or objectivity, as postulated by the machine heuristic [27], might predominate here. Moreover, even if the author is noticed, the person reporting about these number- and data-driven topics might not be of much interest to the reader, which could explain the relatively minor differences found so far. However, even when using a more complex and detailed topic, Lermann Henestrosa et al. [28] found no differences concerning perceived credibility and trustworthiness between an AI and a human author. They discovered at the same time that the AI author was perceived to be less anthropomorphic and intelligent.

1.2. Algorithm Aversion

Evidence shows that people have specific expectations toward AI and algorithms in particular contexts. The word-of-machine-effect describes the belief that AI recommenders are more competent than human recommenders in utilitarian vs. hedonistic realms [29]. In addition, the machine heuristic is a cognitive shortcut when ascribing accuracy or lack of bias to an algorithm when performing certain tasks, for instance, a job in online transactions [27,30]. In line with these findings, algorithm aversion describes the consumers' preference for a human when a task is subjective by nature [31] or concerned with moral decisions because machines are thought to lack a mind and emotions [32]. AI has also been perceived as less competent in giving advice for addressing societal challenges [33]. Applied to AI authorship, Tandoc, Yao, and Wu [23] found a decrease in source and message credibility when the AI was perceived to write non-objectively. In another study, message credibility decreased for both the human and the AI author when the information was presented evaluatively vs. neutrally [28]. With the expanding applicational possibilities of ATGs allowing for the generation of human-sounding text to any possible topic, more research on the perception of AI authorship in different contexts is necessary.

1.3. Surveys on AI and ATG Perception

An online survey by a local initiative for the media and digital scene in Hamburg, Germany, revealed that in 2018, 49% of the respondents were skeptical toward automated news and robot journalism, and 28% considered it "bad", while 20% considered it to depend on the topic [34]. Moreover, in a follow-up survey in 2019, 77% of the respondents demanded that automatically produced content be recognizable as such, while only 39% could distinguish between an actual AI-written text and a human-written one [35]. Interestingly, the wording of the questions in this survey suggested that the prevalence of AI-written texts would be realized only in the future.

A survey among American adults in 2022 revealed a general awareness of the public toward AI in daily life, but only three in ten identified all uses of AI provided in the survey correctly [36]. Another representative survey among the German population investigating the general beliefs and attitudes toward algorithms revealed in 2018 that 45% of the respondents could not indicate what an algorithm is. This knowledge gap was accompanied by skepticism toward algorithms, with 79% of the respondents indicating that they preferred human decisions [37]. Also, different applicational fields of algorithms were not known to a majority but became better known in 2022 [15]. In a recent replication, the authors found evidence for a connection between familiarity and acceptance of automized decisions, with decisions being considered more acceptable and the respondents being more familiar with

the potential field of application [15]. The term algorithm can stand for a simple mathematical function or a highly complex algorithm for data encryption. However, specific knowledge about certain applications and their underlying technology is often irrelevant to users. But, as ATG is now dominating public debate, it is necessary to investigate what people currently think about this specific AI.

1.4. The Current Research

According to the most prominent theory for predicting the acceptance of technology, the technology acceptance model (TAM) [9,38], the adoption of a specific technology is primarily determined by users' performance expectancy and effort expectancy, which influence the attitude toward using it and finally the intention and actual usage of technology. In addition, the evaluation of automatically produced content might depend on participants' attitudes toward AI in general. Darda et al. [39] found that a positive attitude toward AI leads to higher ratings for both automated and human-generated content. Furthermore, discussions around AI fields like automated driving or AI in healthcare were not based on tools suddenly accessible to everyone. In these areas, the focus lies more on the decisions made by AI rather than on the underlying technology that leads to them. This is problematic with tools like ChatGPT, where the information provided will probably be judged based on the user's beliefs about the perceived sources, for instance. However, systematic surveys about people's beliefs, experience, or knowledge concerning specific AI fields are rare, with the majority dealing with perceptions of AI in general or in the medical context [40–43] and surveys even leaving out the specific field of ATG entirely [44]. To the best of our knowledge, there is no current investigation specifically on people's beliefs about ATG or their concepts about its function, responsibilities, or data sources.

In view of these considerations and research gaps, we posed the following research questions: What attitudes, perceptions, and knowledge does the German population have toward ATG? Have these attitudes, perceptions, and knowledge changed over time, specifically since the release of ChatGPT as a critical event? Additionally, we exploratively investigated people's behavioral intentions to consume ATG by using several predictors as suggested by the TAM.

This design made observing a potential change in the data over time possible. Both surveys asked questions about attitude toward ATG, while the second survey included additional questions about ChatGPT to take this event into account as a potential influencing factor. Finally, with its representation of different ages, genders, and educational levels, this study aimed to shed light on differences in the population concerning these subgroups.

2. Methods

2.1. Sample

The study was conducted in accordance with the guidelines of the Local Ethics Committee of the Leibniz-Institut für Wissensmedien, which approved the study design and methods (Approval number: LEK 2023/022). Written informed consent was obtained from all participants involved in the study. Participants were invited to complete the online survey via the online market research platform Mingle in March 2022 (Study 1) and in June 2023 (Study 2). To assure representativeness in terms of age, gender, and education among the German population over 18 years old, quotas were defined in advance. Responses to all questions were voluntary, but participants were only included in the analyses when they had finished the entire survey. Therefore, exclusion criteria were only premature dropout and missing consent to the use of the data. The survey took 10–15 min in the first and 15–20 min in the second census. Each participation was compensated within Mingle's internal reward system.

The Pearson's correlation coefficient with the final sample size of n_1 = 1028 and n_2 = 1013 participants was sufficient to detect correlational effects of r = 0.088 with 80% power (alpha = 0.05, two-tailed), according to sensitivity power analysis (G*Power). In other words, correlations greater than r = 0.088 could be reliably detected.

The participants in Study 1 were on average M_1 = 46.90 (SD_1 = 15.28) years old (range = 18–73 years). The participants in Study 2 had a mean age of M_2 = 45.58 (SD_2 = 14.27) years (range = 18–69). Table 1 shows the absolute and relative distributions by survey for gender, education, and age.

Table 1. Absolute and relative (in percent) numbers of participants in Studies 1 and 2 by gender, education level, and age group. The German education system differentiates between qualifications after the 9th (low), 10th (middle), and 12th or 13th grade (high). Respondents indicating having no degree (n_1 = 13, n_2 = 4) are included in the low level.

	Study 1		Study 2	
	n	**%**	***n***	**%**
Gender				
male	527	51.26	516	50.94
female	499	48.54	495	48.86
diverse	2	0.19	2	0.20
Educational level				
low	346	33.66	270	26.65
middle	324	31.52	349	34.45
high	358	34.82	394	38.89
Age group				
18–29	184	17.90	188	18.56
30–39	182	17.70	188	18.56
40–49	175	17.02	188	18.56
50–59	232	22.57	248	24.48
>60	255	24.81	201	19.84

2.2. Measures and Procedure

In the following paragraphs, we describe the measures and procedure of both surveys, as Study 2 was conducted in the same way as Study 1 apart from several additional questions concerning ChatGPT. After giving informed consent and their initial screening with respect to age, gender, and education, respondents were redirected to the survey platform Qualtrics (Provo, UT, USA).

First, self-assessed knowledge about AI in general was captured with a single item from 1 = no knowledge about AI to 5 = comprehensive knowledge about AI.

Afterwards, participants were briefly introduced to the topic and were presented with a general definition of AI, followed by a section about AI in general.

General attitudes toward AI were assessed by using a 20-item instrument [45]. The measure comprised 12 positively (e.g., "There are many useful applications of AI") and eight negatively phrased sentences (e.g., "I think AI systems make many mistakes"). These were worded to express a general attitude toward AI systems mainly in society and in the work context. The items were measured on 5-point Likert-scales from 1 = absolutely disagree to 5 = absolutely agree.

The belief in the machine heuristic [27] was used to further measure participants' assessment of AI. Participants were asked for their degree of agreement, from 1 = absolutely disagree to 5 = absolutely agree, with four adjectives for an AI when performing a task ("unbiased", "error-free", "objective", and "accurate").

Then, participants were asked if and how often they used different speech-based applications (e.g., voice-based assistants, chatbots, translation systems) in order to examine experience with AI-based writing- and voice-software among the population (5-point scale from 1 = never to 5 = constantly). Moreover, we asked if and how often they used different media types (e.g., TV, radio, magazines) to obtain information about scientific topics. To both questions "ChatGPT" was added as an option in Study 2, which served to analyze further the subgroups with and without ChatGPT experience. Only participants who indicated having used ChatGPT were presented with the attitude toward ChatGPT scale.

Subsequently, we assessed participants' experience with ATG. Participants were asked if they had ever heard about the fact that AI is able to write texts (Heard about ATG) and if they had ever consciously read an AI-written text (Read an AI generated text), both on 5-point single items from 1 = never to 5 = constantly. If participants indicated with at least item 2 = seldom to have read a text written by AI, they were redirected to an open response field and were asked to state the type of text(s) they had read so far.

At this point in Study 2, a knowledge test about ATG followed. The test covered 15 partly adapted [46] statements, for which participants had to decide whether they were true, false, or if they didn't know (e.g., "Humans can still easily recognize AI-generated speech as artificial speech").

In both surveys, a short description and definition of ATG and automated journalism followed to assure that every participant had at least a basic understanding of the subject matter. To keep it as simple as possible, we consistently used the phrases "AI-written text" or "AI-generated text".

The definition was followed by a set of self-generated statements to examine people's conceptions about ATG. The scales referred to the mode of ATG's function (ATG functionality), the source of the automatically written texts' content (data sources), the extent of control participants believed a human has over an AI-written text (human control), and who they believed was responsible for the content (content responsibility). Participants rated their perceived likelihood of each item (5-point scales from 1 = not at all to 5 = for certain).

People's understanding about ATG functionality was assessed by four statements in Study 1 (e.g., "The AI uses existing words and texts and reassembles them"). In Study 2, the item "The AI calculates the word that is most likely to follow next" was added, as this applies to the LLM underlying ChatGPT. Respondents' belief about the data sources was assessed by five items in Study 1 (e.g., "The AI produces the content itself, without human intervention"). In Study 2, the item "The AI has been trained with certain content, which it then draws on" was added, as this is a more precise description of ChatGPT's functionality. To assess people's belief about human control, we presented three items in Study 1 (e.g., "The human sees the final product and edits it if necessary"). Again, in Study 2, the item "The end product is only indirectly controlled by humans (via built-in rules)" was added. Then, participants were presented with eight entities (e.g., "programmer" or "AI itself") for which they had to indicate the likelihood that one or the other was responsible for the produced content.

The concepts section was followed by five adapted subscales from the UTAUT-instrument (Unified Theory of Acceptance and Use of Technology) [9] to measure specific attitudes toward AI-written texts. All items were measured on 5-point Likert scales from 1 = absolutely disagree to 5 = absolutely agree.

Three items were presented concerning performance expectancy (e.g., "I would find AI-written texts useful"), three items concerning effort expectancy (e.g., "I think AI-written texts are clear and understandable"), four items concerning participants' attitude toward using ATG (AT; e.g., "AI-written texts would make information retrieval more interesting"), three items concerning anxiety (e.g., "AI-written texts are somewhat intimidating to me"), and three items concerning behavioral intentions to consume ATG (e.g., "I intend to read AI-written texts in the future"). To assess participants' attitude toward what an AI should be permitted to write (permission to write like a human), we added four self-created items (e.g., "AI should be allowed to write about the same topics humans do").

At the end of the specific attitudes block, participants were asked how likely (5-point scales from 1 = not at all to 5 = for certain) they would be to read an AI-written text on 18 different news media topics (e.g., politics, society, or weather forecasts). Afterwards, the identical list of topics was presented again, asking participants to indicate if they could choose freely by whom they would prefer to read about each topic ("preferably by a human being", "no preference", "preferably by an AI").

In Study 2, participants who had experience with ChatGPT were asked to indicate their agreement to 16 statements addressing their attitude toward ChatGPT (e.g., "I am

satisfied with ChatGPT's answers"; 5-point Likert scale from 1 = absolutely disagree to 5 = absolutely agree).

Finally, independently of their prior experience, all participants in Study 2 were presented with a definition of ChatGPT and were afterwards asked about their lay attitude toward ChatGPT and ATG by indicating their agreement with nine statements (e.g., "I'm optimistic about the impact of automated text generation (e.g., ChatGPT) on society"; 5-point Likert scale from 1 = absolutely disagree to 5 = absolutely agree).

3. Results

3.1. General Attitudes toward AI

Participants' answers on the item concerning their self-assessed knowledge about AI are displayed separated by gender and study (Figure 1) and by education and study (Figure 2). Table 2 shows the means, standard deviations, and Cronbach alpha values for all single items and scales by study. In addition, exploratory *t*-tests for independent samples were conducted to compare the two time points. Furthermore, the relationships between the variables in Study 2 are depicted by Figure 3 showing the correlations between all Likert-type variables and the knowledge test (for means and standard deviations of all variables separated by age groups, see Supplementary Table S1).

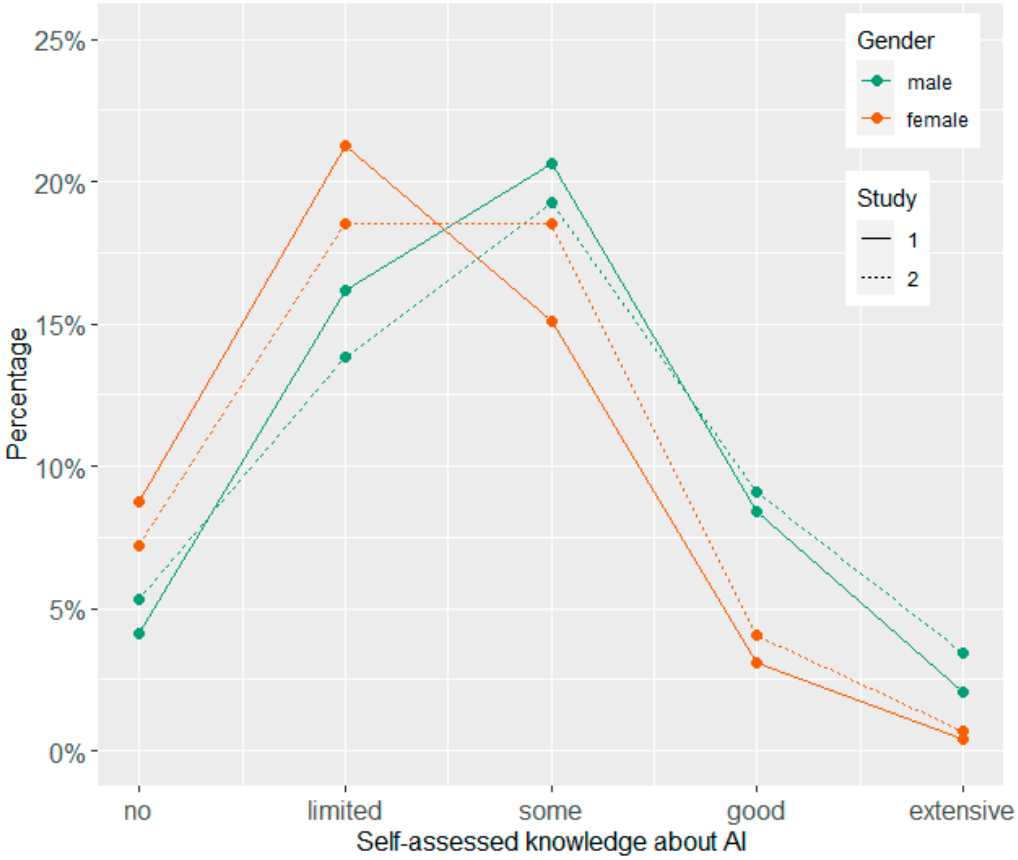


Figure 1. Self-assessed knowledge about AI by gender and study survey.

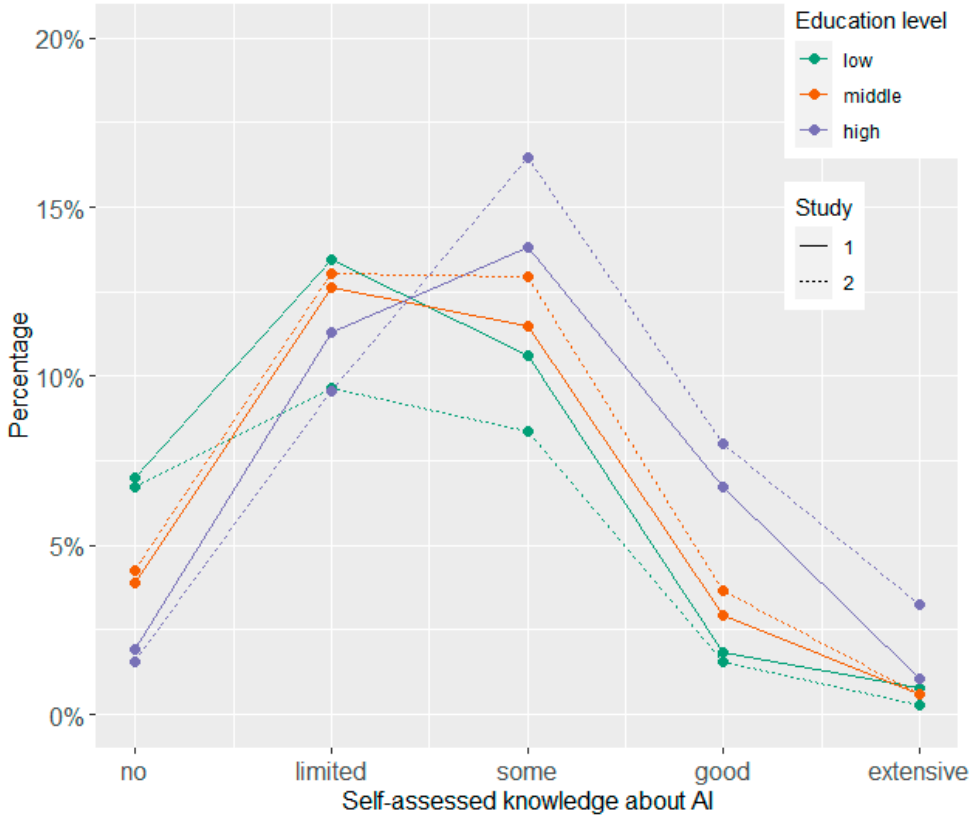


Figure 2. Self-assessed knowledge about AI by education level and study survey.

Table 2. Mean, standard deviation, and Cronbach alpha for all variables (number of items in brackets) by study and differences in means (ΔM) with Cohen's d, tested using exploratory unpaired *t*-tests; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. All variables except for knowledge were measured on 5-point scales.

	Study 1 (*n* = 1028)		**Study 2 (*n* = 1013)**				
Variable	***M***	***SD***	***M***	***SD***	**α**	**Δ*M***	***d***
Self-assessed knowledge (1)	2.53	0.94	2.64	1.00	-	0.11 *	0.11
General attitudes (20)	3.16	0.64	3.00	0.69	0.92	−0.16 ***	0.24
Machine heuristic (4)	3.46	0.73	3.30	0.80	0.79	−0.16 ***	0.21
Knowledge (15)	-	-	5.88	3.37	-	-	-
Performance expectancy (3)	2.85	0.83	2.96	0.90	0.81	0.11 **	0.12
Effort expectancy (3)	3.21	0.74	3.31	0.79	0.71	0.09 **	0.12
Attitude twd using ATG (4)	3.01	0.80	3.03	0.87	0.79	0.02	-
Anxiety (3)	2.69	0.83	2.80	0.86	0.71	0.11 **	0.13
Behavioral intentions to consume ATG (3)	2.77	0.83	2.78	0.92	0.67	0.01	-
Permission (4)	2.67	0.72	2.75	0.73	0.62	0.08 *	0.11
Attitude twd ChatGPT (16)	-	-	3.16	0.61	0.89	-	-
Lay attitude twd ChatGPT & ATG (9)	-	-	2.88	0.73	0.82	-	-

Due to the poor internal consistency of the scale, which indicates that the items covered different aspects, participants' answers to each of the four self-created items capturing the permission to write like a human are presented separately. The means and standard deviations by study were as follows: $M_1 = 2.98$ ($SD_1 = 1.06$) and $M_2 = 3.05$ ($SD_2 = 1.09$) for "AI should be allowed to write about the same topics as humans", $M_1 = 3.70$ ($SD_1 = 0.96$) and $M_2 = 3.71$ ($SD_2 = 1.00$) for "AI should present pure facts" (reversely scored in the scale), $M_1 = 2.59$ ($SD_1 = 1.08$) and $M_2 = 2.67$ ($SD_2 = 1.12$) for "AI may express an opinion

in its texts", and $M_1 = 2.82$ ($SD_1 = 1.07$) and $M_2 = 2.99$ ($SD_2 = 1.14$) for "The AI may write emotional texts".

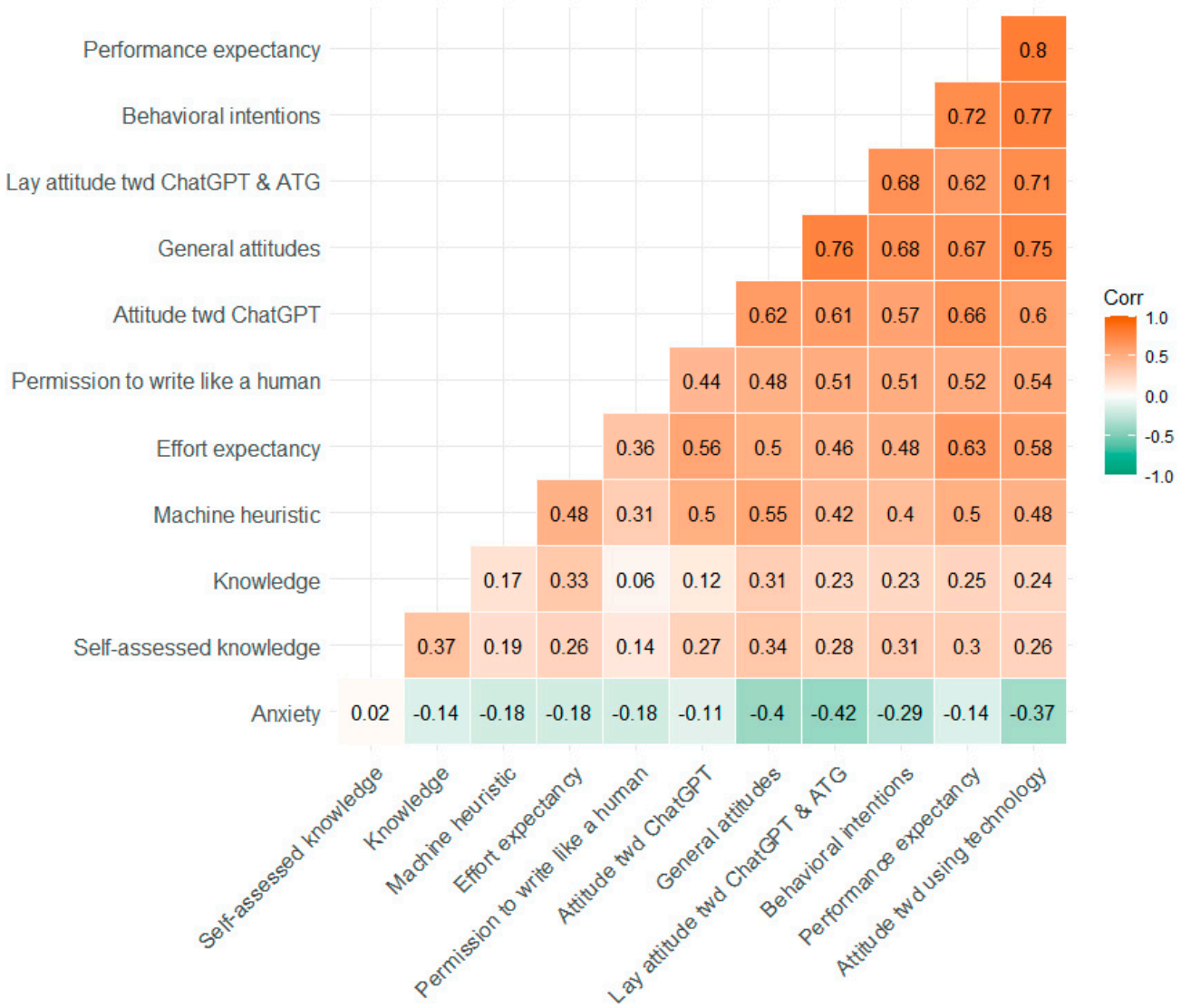


Figure 3. Correlation matrix for all Likert-type variables and the knowledge test in Study 2. The responses of the knowledge test were re-coded for the purpose of the analyses, i.e., the answer option "don't know" was coded as 0 = false response.

3.2. Experience with ATG

The responses to the questions of whether participants had ever heard of ATG and whether participants had ever read a text written by AI can be seen in Table 3. In addition, the answers specifically regarding ChatGPT use are displayed in this table. A total number of n = 408 participants in Study 2 (40.28% of respondents) indicated having used ChatGPT before and were thus later forwarded to the attitudes toward ChatGPT questionnaire (see below). For distributions of the scales separated for participants who indicated having or not having used ChatGPT, see Figure 4.

Table 3. Relative frequencies (in percent) of answers on the items regarding ATG experience and ChatGPT use by study. * Answer options seldom, occasionally, often, and constantly were merged.

Item	Answer Option	Percentage (%)	
		Study 1 (n = 1028)	**Study 2 (n = 1013)**
Heard about ATG	Never	32.49	16.09
	Seldom	17.41	10.46
	Occasionally	32.68	32.77
	Often	15.37	27.44
	Constantly	2.04	13.23
Read an AI-generated text	Never	56.52	48.86
	Seldom	21.01	17.28
	Occasionally	16.63	22.01
	Often	5.06	8.19
	Constantly	0.78	3.65
General ChatGPT use	Never	-	64.76
	At least seldom *	-	35.24
ChatGPT use for scientific information	Never	-	65.25
	At least seldom *	-	34.75

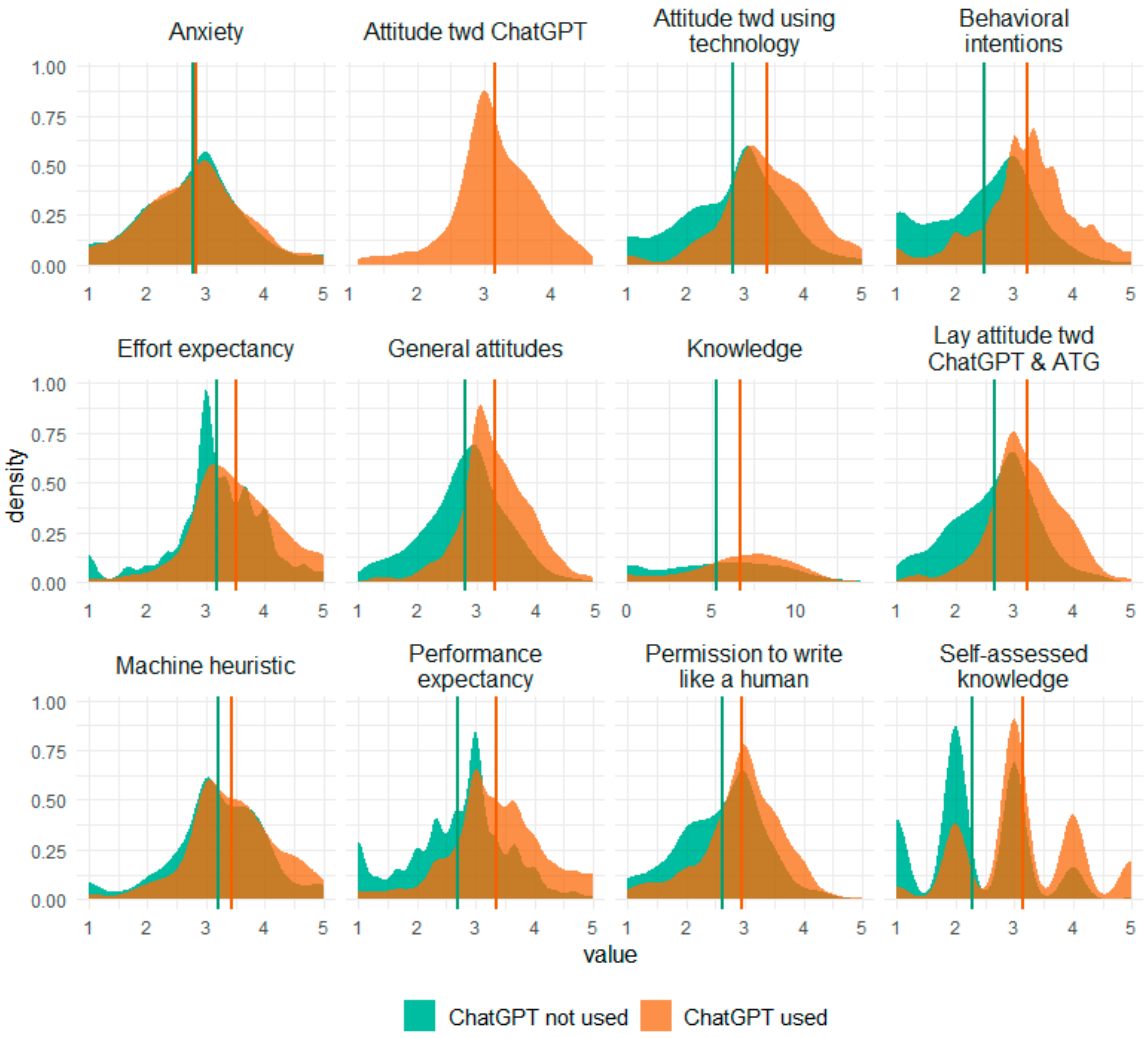


Figure 4. Density distributions of all variables for the subgroups of ChatGPT use in Study 2. Vertical lines represent the means within the subgroups; n = 605 for "ChatGPT not used" and n = 408 for "ChatGPT used".

3.3. Knowledge Test

Concerning the knowledge test in Study 2, 120 (11.85%) respondents did not answer any question correctly while only two (0.002%) reached to answer 14 statements correctly (see Table 2 for mean and standard deviation). For the 15 statements of the knowledge test and the distribution of participants' responses on each statement, see Table 4.

Table 4. Relative frequencies (in percent) of the correctness of the answers on the knowledge test and correctness of each statement (True/False) in Study 2 (n = 1013).

			Percentage (%) of Answers		
Statement		**Correctness of Statement**	**Correct**	**Incorrect**	**Unknowing**
1.	Pupils can have their homework created with the help of speech-generating AI	True	59.13	11.55	29.32
2.	There are different types of models used for ATG	True	57.16	6.81	36.03
3.	ChatGPT has been trained with millions of texts from the web, social media, online forums, newspaper articles and books	True	56.37	6.32	37.31
4.	The statements of language-generating AI are always correct	False	56.27	11.45	32.28
5.	Access to language-generating AI is reserved for certain groups of people (e.g., scientists)	False	52.52	12.14	35.34
6.	Speech-generating AI responses may be biased (e.g., racially) based on the data they were trained on	True	46.40	12.83	40.77
7.	A chatbot can answer the question 'Will it rain tomorrow?' correctly with a high probability	True	45.01	21.42	33.56
8.	AI language models (e.g., ChatGPT) calculate for their answers which word is most likely to come next	True	44.52	10.56	44.92
9.	Humans can still easily recognize AI-generated speech as artificial speech	False	43.53	23.30	33.17
10.	AI language models can intentionally lie and spread false information	False	28.43	31.00	40.57
11.	Humans can answer questions about a read text better than AI systems	False	26.16	33.17	40.67
12.	AI language models (e.g., chatbots) can give good answers because they have learned to understand language like a human	False	20.53	47.29	32.18
13.	The automatic generation of texts has been used in journalism for over 10 years	True	20.24	20.42	59.33
14.	Texts created by AI must be legally marked as such	False	17.67	39.59	42.74
15.	The quality of the texts created by AI depends only on the training data set used	False	14.41	49.06	36.53

3.4. Concepts

Figures 5–8 show participants' perceived probabilities toward each item concerning the concepts. Relative answers to each belief about how ATG could potentially work (Figure 5), where the content could come from (Figure 6), how much control the human could have in the process (Figure 7), and who could be responsible for the content (Figure 8) are depicted by study.

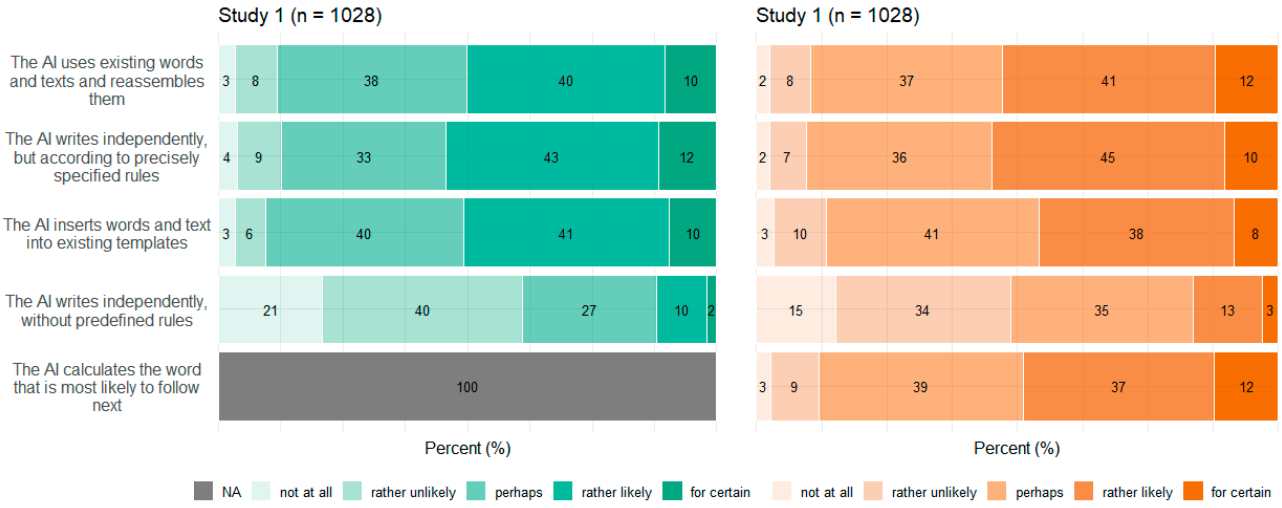


Figure 5. Relative frequencies of participants' perceived probability for each item regarding ATG functionality by study.

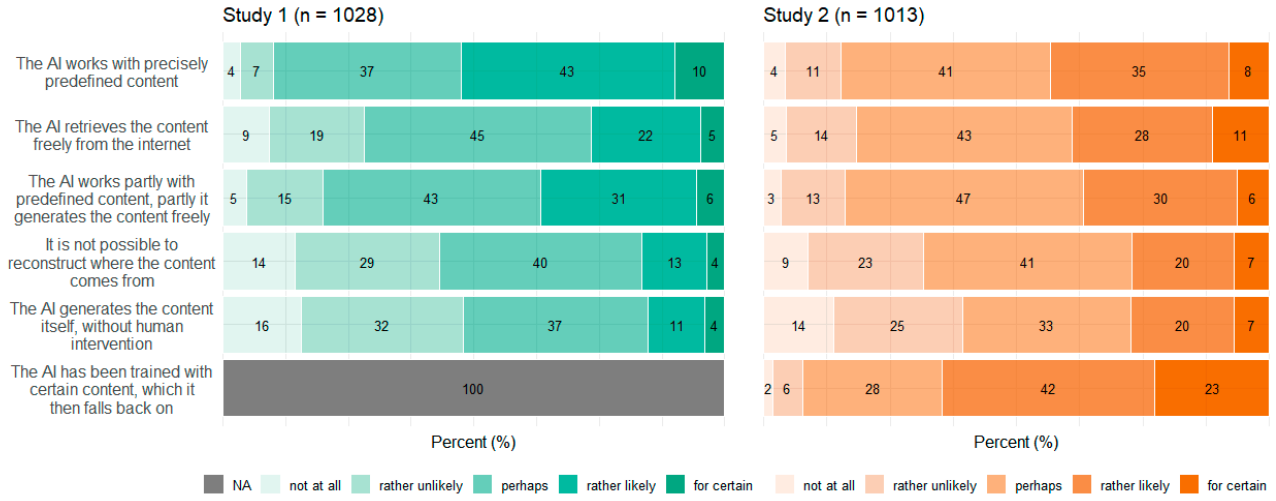


Figure 6. Relative frequencies of participants' perceived probability for each item regarding data sources by study.

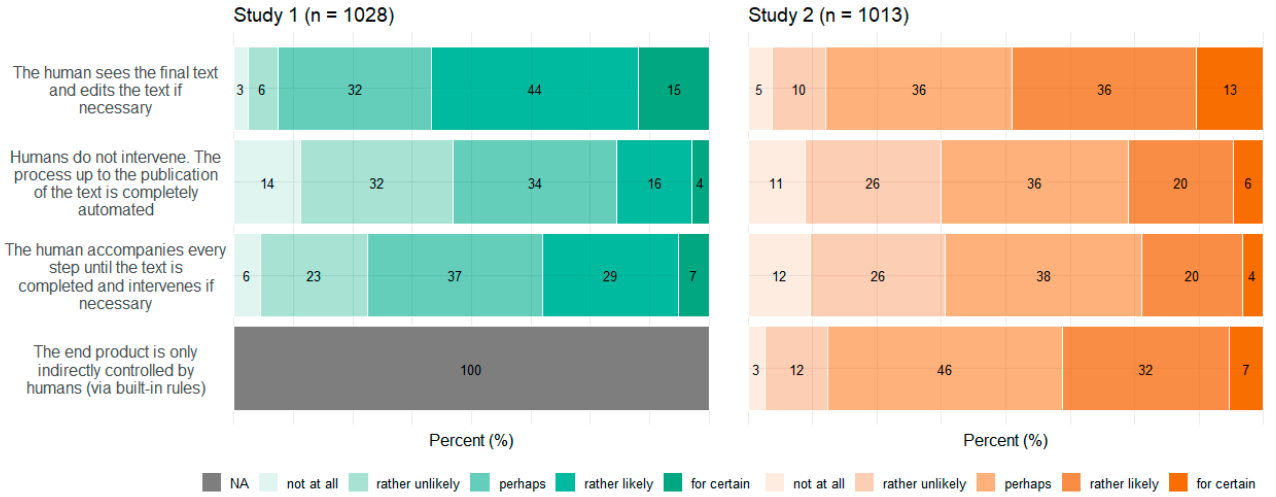


Figure 7. Relative frequencies (percentage) of participants' perceived probability for each item regarding human control by study.

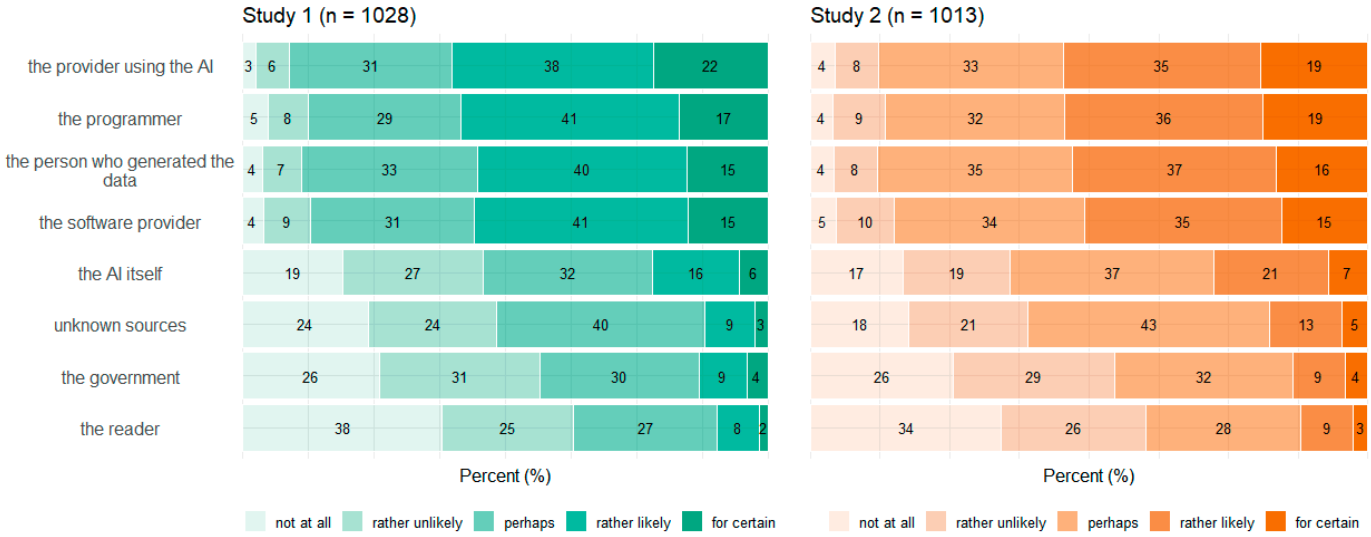


Figure 8. Relative frequencies (percentage) of participants' perceived probability for each item regarding content responsibility by study.

3.5. Intention to Read AI-Written Texts concerning Journalistic Topics

For participants' detailed answers on the question "How likely would you be to read AI-written texts on the following topics?" separated by study, see Supplementary Figure S1. Regarding all 18 topics, the proportion of participants answering "perhaps" was between 30–40% for both studies. The option to read an AI written text "for certain" was chosen in equally small proportions in both studies, with the highest values for product descriptions and weather forecasts. The proportion of participants indicating to be "not at all" willing to read an AI-generated text to the presented topics reached between 8–30% in Study 1 and 12–26% in Study 2, with the highest proportion of rejection for the topic "opinion" in both time points.

Concerning the question "If you had a choice, who would you rather be informed by about the following topics?", detailed response distribution across the 18 topics is depicted in Supplementary Figure S2. At both time points and regarding all topics, the percentage of people who chose "preferably by an AI" did not reach a majority. The proportion of participants preferring an AI author was between 3–21% in Study 1 and 5–20% in Study 2, depending on the topic. On some topics, respondents expressed a clear tendency to prefer a human author (e.g., opinion, health, politics), whereas on some topics the option "no preference" was selected more frequently (e.g., sports reports, advertisements, stock reports).

3.6. Specific Attitudes toward ATG

In Study 2, participants who indicated having used ChatGPT were asked about their attitudes and their experience with this tool. The answer distributions regarding each item are depicted in Table 5. Furthermore, all participants indicated their attitudes toward ATG and ChatGPT on a more general level (Table 6).

Table 5. Relative frequencies (in percent) of participants' agreement to each item regarding attitude toward ChatGPT in Study 2 (n = 408).

		Percentage (%) of Answers				
Item		**Absolutely Disagree**	**Rather Disagree**	**Undecided**	**Rather Agree**	**Absolutely Agree**
1.	When I'm unsure with something I would rather trust ChatGPT than me.	17.40	25.49	34.31	18.38	4.41
2.	The answers ChatGPT provides are as good as the answers a highly competent person would give.	8.82	19.12	42.16	22.79	7.11
3.	I will recommend ChatGPT.	5.88	9.80	41.42	29.41	13.48
4.	I will continue to use ChatGPT.	3.68	4.90	40.20	33.82	17.40
5.	I trust ChatGPT.	7.11	14.71	45.83	24.51	7.84
6.	I still prefer receiving answers and texts from a human.	2.45	11.03	42.89	27.45	16.18
7.	I like ChatGPT.	4.90	6.37	46.81	28.92	12.99
8.	I know how to use ChatGPT to get the results I need.	5.88	13.97	39.95	30.88	9.31
9.	I doubt the answers ChatGPT gives me.	5.15	21.57	47.79	18.87	6.62
10.	I can rely on ChatGPT's answers when it comes to decisions.	8.33	15.44	44.61	24.26	7.35
11.	I believe in ChatGPT's answers even if I can't be sure they're right.	8.82	21.32	43.63	21.57	4.66
12.	I'm satisfied with ChatGPT's answers.	4.41	8.33	46.08	32.11	9.07
13.	ChatGPT uses appropriate methods to generate its answers.	3.19	6.62	41.91	37.99	10.29
14.	ChatGPT provides me all information I need.	6.37	11.76	45.83	27.70	8.33
15.	ChatGPT's answers meet my expectations.	4.17	11.03	42.16	33.82	8.82
16.	ChatGPT's answers are unusable.	18.63	29.41	34.07	14.71	3.19

Table 6. Relative frequencies (in percent) of participants' agreement to each item regarding lay attitude toward ATG and ChatGPT in Study 2 (n = 1013).

Item		Percentage (%) of Answers				
		Absolutely Disagree	Rather Disagree	Undecided	Rather Agree	Absolutely Agree
1.	There should be a labelling requirement for AI-generated texts.	3.26	4.24	20.04	28.13	44.32
2.	Policymakers should make precise rules on where automated text generation may be applied.	5.53	8.09	28.33	31.00	27.05
3.	I feel uneasy with ChatGPT.	13.03	18.26	34.55	19.64	14.51
4.	I intend to try ChatGPT.	19.25	14.31	30.21	22.31	13.92
5.	I find ChatGPT dangerous.	11.35	19.55	38.50	18.85	11.75
6.	I understand how ChatGPT works.	6.42	12.34	40.28	30.90	10.07
7.	ChatGPT should be used for scientific information, too.	11.85	14.02	40.08	24.68	9.38
8.	ChatGPT should be prohibited.	26.65	26.36	28.83	9.97	8.19
9.	I'm optimistic about the impact of automated text generation (e.g., ChatGPT) on society.	13.43	17.57	42.65	20.63	5.73

3.7. Exploratory Analyses

In both studies, we measured participants' intentions to read AI-written texts and to use ATG technology. Due to the release of ChatGPT in November 2022, we were able to ask more specifically for attitudes toward ChatGPT and ATG in Study 2. Therefore, we conducted two explorative multiple regression analyses to predict people's behavioral intentions to consume ATG. Figure 9 depicts Q-Q plots for evaluating residual normality in the models for both studies. The plot shows that the residuals follow the theoretical quantiles of the normal distribution well around the mean, with some deviation at the tails of the distribution. This deviation is expected (and commonly seen) since the theoretical normal distribution ranges from minus infinity to infinity, which is, of course, not true for our measure.

We adopted the common supervised machine learning principle k-fold cross-validation, which helps to estimate the predictive accuracy of a regression model. Using cross-validation, the data set is divided into a training (used for model training) and a test set (also hold-out set, used to evaluate the model performance). The purpose of this approach is not to fit the model to the entire sample but only to a part of it (training set) and thus to test whether the model can be generalized to the unseen hold-out set. Furthermore, the training set is divided into k equal-sized folds on which the statistical model is iteratively developed and fitted, leaving each fold out in turn. This process is repeated k times, with each fold being used as the validation set once. Therefore, cross-validation is a robust and reliable method to reduce the risk of model overfitting and serves the generalizability of the regression results to unseen data [47–49].

According to the TAM [9], the intention to use a technology is determined by people's attitude toward using it, which is in turn influenced by the performance expectancy and effort expectancy. Therefore, we added these variables to the models. We also aimed

to investigate the predictive contribution of several other variables, such as the general attitude toward AI or participants' prior experience with ChatGPT, as relationships between these variables were found before [39,50].

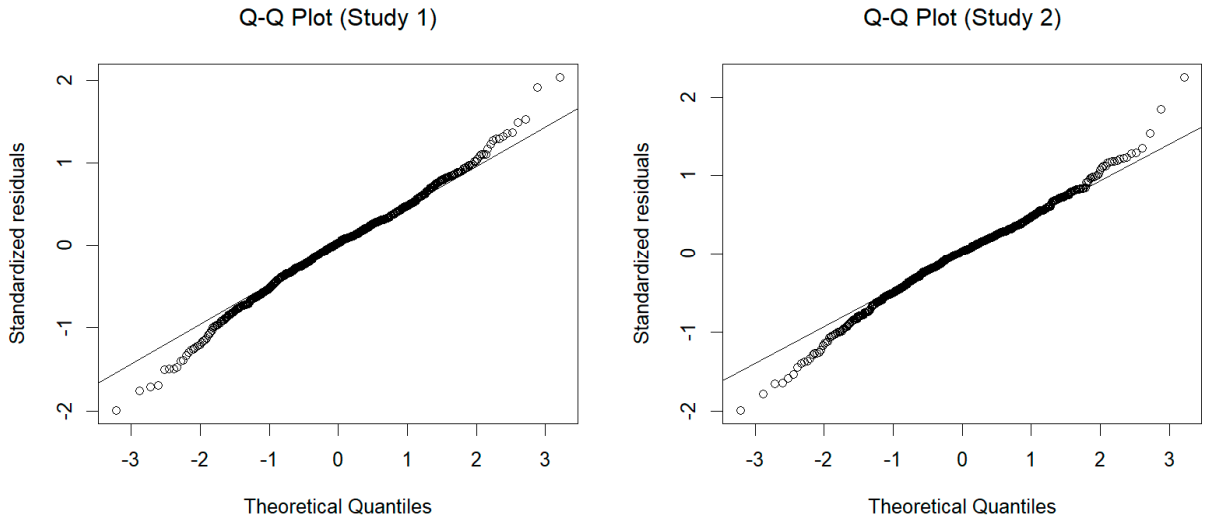


Figure 9. Q-Q plots of residuals to assess normality in the regression models predicting behavioral intention to consume ATG for both studies.

Each survey data set was split into a training (75%) and a hold-out set (25%). A regression model using k = 5-fold cross-validation was used in each training set. This process involved training the model in four subsets and evaluating it in the remaining subset in a rotating fashion. Finally, the performance on the fitted model was assessed in the hold-out sample. The predictors of the respective models as well as their corresponding coefficients can be seen in Table 4 and are illustrated in Figure 10. Results indicated that both models explained a substantial proportion of variance, with $R^2_1 = 0.66$ and $R^2_2 = 0.66$. Together with the small average magnitudes of errors, $RMSE_1 = 0.48$ and $RMSE_2 = 0.53$, and mean absolute errors, $MAE_1 = 0.37$ and $MAE_2 = 0.43$, the fitted models were highly accurate in their predictions. In both models, as expected, performance expectancy and attitude toward using ATG significantly contributed to predicting behavioral intentions, but effort expectancy did not. Moreover, in Study 2, the added predictor lay attitude toward ChatGPT and ATG was a significant predictor (see Table 7).

Table 7. Coefficients of the multiple regression analyses with cross-validation for predicting behavioral intentions to consume ATG in Studies 1 and 2. Heard about ATG, read an AI-generated text, and ChatGPT use were included in the models as dummy variables with 0 (not heard, not read, and no usage) serving as the reference group.

	Study 1			**Study 2**		
	b	***SE***	***p***	***b***	***SE***	***p***
Self-assessed knowledge	0.05	0.02	0.031	0.02	0.02	0.409
General attitudes	0.13	0.04	0.003	0.13	0.05	0.014
Machine heuristic	−0.06	0.03	0.050	−0.05	0.03	0.084
Heard about ATG	0.07	0.05	0.132	0.04	0.06	0.480
Read an AI-generated text	0.03	0.04	0.480	0.13	0.05	0.008
Performance expectancy	0.17	0.04	<0.001	0.25	0.04	<0.001
Effort expectancy	−0.02	0.03	0.544	−0.03	0.03	0.438
Anxiety	−0.02	0.03	0.517	−0.04	0.03	0.177
Attitude twd using ATG	0.62	0.04	<0.001	0.41	0.05	<0.001

Table 7. *Cont.*

	Study 1			Study 2		
	b	***SE***	***p***	***b***	***SE***	***p***
Knowledge	-	-	-	−0.01	0.01	0.315
Lay attitude twd ChatGPT and ATG	-	-	-	0.20	0.05	<0.001
ChatGPT use	-	-	-	0.07	0.05	0.130

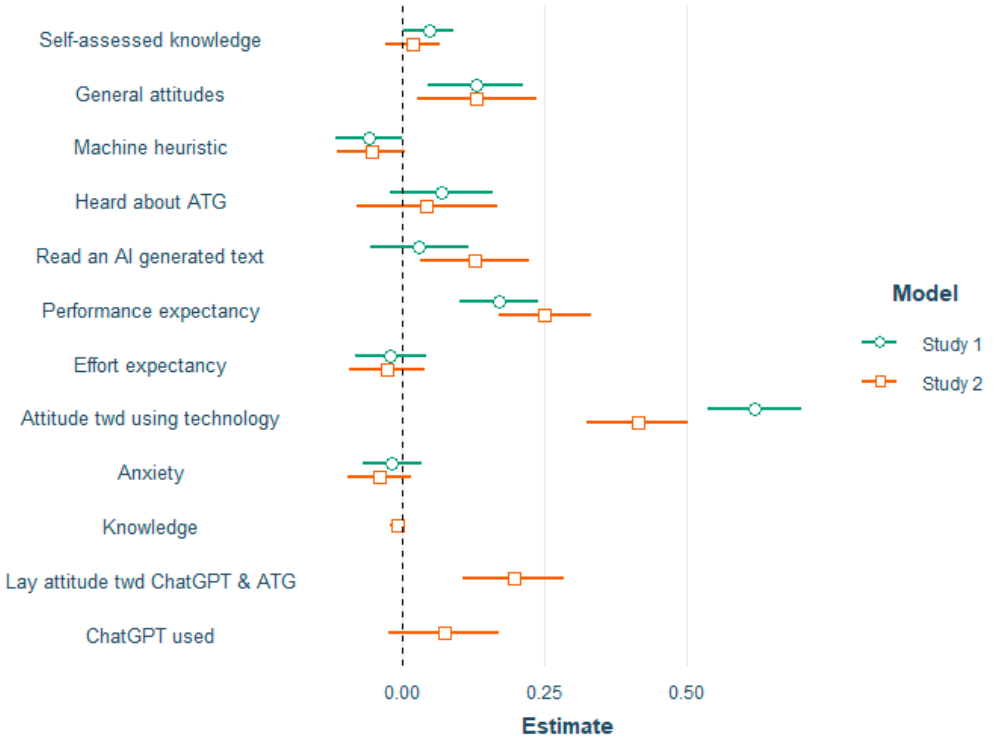


Figure 10. Coefficients of the predictors of behavioral intention to consume ATG for both models (Study 1 & Study 2). Whiskers represent the 95% CI.

4. Discussion

What society expects from technological development, which information people may need to understand and adopt innovations, or whether the attitudes tend to be positive or negative should be investigated parallel to the developmental process. Awareness of people's perceptions and ideas regarding technologies that are ever more present is essential for a meaningful debate about developments and adequate information for laypeople. However, discussions such as those that have been held publicly about deep fakes or facial recognition technology are missing regarding ATG, as public perception is currently largely shaped by one new tool. Even if many people are currently not aware of where ATG is already being used or cannot distinguish between human and AI-written texts, research into attitudes and possible concerns is essential, as these influence trust in and the handling of texts authored by AI. Moreover, knowledge gaps need to be revealed to target misconceptions. We addressed this research gap by investigating the German population's current beliefs, concepts, and attitudes toward ATG. Data from two representative surveys conducted in 2022, before the release of ChatGPT, and 2023, after the sudden media focus on NLG developments, were collected to gain insights into the current state and potential changes over time.

4.1. Public Awareness of ATG

Without a doubt, the hype surrounding ChatGPT has drawn attention to this field of AI. However, a survey among the U.S. population finding that 42% of Americans had in March 2023 heard nothing at all about ChatGPT revealed that this did not reach all sections of the population equally quickly [36]. While a third of respondents in Study 1 presented here indicated never having heard that AI can write texts, this proportion decreased substantially to 16% in Study 2. In contrast, the proportion of people indicating never having read an AI-written text barely dropped from 56% to 49% between the two polls. Apparently, the coverage of ChatGPT has led people to become more concerned with this technology. Though we cannot retrospectively measure participants' actual experience with ATG, the evidence still reflects a remarkable lack of awareness of the presence of ATG. Automatically produced content has been present in automated journalism for over a decade (unknown to a majority, as the knowledge test indicated), and automatic product descriptions in online shops are also not new phenomena. Similarly, a German survey by DIW Berlin revealed that many people are unaware of AI in their work contexts: when indirectly asked about AI at work, nearly twice as many respondents indicated working with AI compared to when they were directly asked [51]. It is important to consider that the criteria for what constituted AI are also shifting. The futuristic, distant, and complex image many associated with AI certainly differs from their image of a simpler translation algorithm. Furthermore, an in-depth understanding of technology is not decisive for people to apply it. Nevertheless, future research should constantly adapt to the given circumstances and accompany technological developments.

4.2. Self-Assessed Knowledge about AI

Concerning self-assessed knowledge about AI, we observed only small shifts from Study 1 to Study 2. However, there were differences on gender and education levels. Men rated their knowledge higher slightly more often than women, while women indicated more frequently than men having limited or no knowledge. In addition, the higher the educational level, the higher the proportion was of participants who indicated some or good knowledge about AI in general. Similar to a survey from Bertelsmann Stiftung [37], we found that people who indicated knowing more about AI in general had a more positive attitude toward it. Moreover, we also found a positive but rather moderate correlation with the performance in the knowledge test. An educational advantage could enable people to more effectively determine how to use technology to their advantage.

4.3. Attitudes toward ATG

We explored significant differences between the two time points on most scales that were surveyed twice. The general attitudes toward AI and the belief in the machine heuristic decreased in Study 2, reflecting a less positive attitude and a slightly diminished belief in the accuracy and objectivity of AI. However, the effect sizes do not allow any conclusion of substantial practical relevance. Rather, this study observed stability across the concepts over time.

Furthermore, high positive correlations occurred among the attitude scales (see Figure 3): The more positive the attitude toward AI, the more favorable was the attitude toward ChatGPT and ATG as well, a relationship also found among a student sample in Arabic countries [52]. However, more intriguing is the positive relationship between the machine heuristic (i.e., the belief that an AI is objective, accurate, neutral, and unbiased) and the attitude scales, as they suggest that predispositions can predict reactions to specific technologies. Furthermore, anxiety was negatively correlated, especially with lay attitude toward ChatGPT and ATG and the general attitudes toward AI, a pattern already found before [53] and in other cultural contexts [52]. More research is needed to address concerns and worries to specifically enable people to deal with ATG appropriately. Tool-specific strengths and weaknesses can get lost in public debates about "generative AI". Educational gaps can either be bridged or widened by ATG technology, particularly when they are pri-

marily used by groups that already benefit from new developments more than others. This underscores the necessity of widespread and transparent information about the potentials and limits of ATG.

4.4. Knowledge Regarding ATG

The knowledge test conducted in Study 2 revealed some detailed insights into potential misconceptions and knowledge gaps. Almost half of the participants incorrectly believed that the quality of AI-created texts depends only on the training data set. That many also believed that language models have learned to understand language like a human. Of course, these two items require a high level of technological understanding. However, the distribution of answers could reflect a misconception about the fundamental function of AI and ATG: It is a popular misunderstanding that artificial and human intelligence work in the same way. Whereas AI developers aim to imitate human intelligence orienting on the mere results, the technological process of reaching these (seemingly) intelligent results can differ significantly from human cognitive or physical processes. The statements correctly answered by a considerable proportion of participants covered more general aspects with no need for intense technical understanding (e.g., "The statements of language-generating AIs are always correct"). Overall, a high proportion of participants responded "don't know" to each statement, reflecting the heterogeneous level of knowledge in the population but also emphasizing the need for more information and explanation about ATG.

4.5. ATG-Related Concepts

Regarding the four concepts relating to ATG, respondents had to assess the likelihood that each statement applied, as the concepts covered different possibilities rather than hard facts. However, within each item of the concepts function of ATG and data source, a high proportion of participants chose the option "perhaps". No clear tendency toward single statements was observed within any of the four concepts. Still, a substantial portion of participants perceived most ideas as "rather likely" or "rather unlikely". Only a small fraction committed to the answer options "for certain" or "not at all". In Study 2, an item was added to three of the concepts, to cover ChatGPT's mechanisms. Concerning function of ATG, the added item "The AI calculates the word that is most likely to follow next" (which describes the function very simply) was perceived to be nearly equally likely as the other options. Only the statement that the AI writes independently was perceived as the least probable option. Concerning data source, participants perceived it to be most likely that the AI would have been trained with certain content, which it then falls back on. Moreover, the items "The AI retrieves the content freely from the internet" as well as "It is not possible to reconstruct where the content comes from" and "The AI generates the content itself" were perceived to be likely more often than in Study 1.

Overall, the high frequency of the answer option "perhaps" along with the result that no statements sticked out with a clear tendency of being favored speak for a great uncertainty concerning the underlying mechanisms of ATG. Similarly, a current survey on ChatGPT found high proportions of indecision, too [54]. Some items also reflected different existing technological approaches of ATG, which do not necessarily contradict each other. It is also unrealistic that users would be comprehensively informed about all the specific underlying technologies. Nevertheless, a basic understanding of the potentials and limitations of ATG is crucial for a realistic assessment of its application and use. Furthermore, with the release of tools available for people with every conceivable level of knowledge, benchmark studies such as the present one can provide information about misconceptions and possible weaknesses of technologies that must be addressed.

4.6. Preferences for Human Authorship

The current study shows that the broad field of topics that falls under "news" has to be examined in a more sophisticated way, since people seem to have topic-specific preferences. Similar to what "algorithm aversion" [31] predicts, respondents are more

likely to read AI-written content about objective and impersonal topics (e.g., traffic news, weather forecasts, or product descriptions). In contrast, participants particularly refuse to read AI texts about genuinely human-centered topics [55] (e.g., society, culture, or politics). When directly asked for author preferences, participants prefer human authorship across all 18 topics, even with slight shifts toward human preference in Study 2. In both surveys, the notable number of participants selecting "perhaps" or "no preference", along with many indicating they have never read AI-written texts, suggests uncertainty or a lack of imagination about what AI-created content entails. The results are remarkable against the background of a tool that aims to completely imitate written human language in all areas. At the same time, in the context of science communication, two studies suggest that readers perceive a human and an AI author as equally credible [28] or only slightly less credible [56]. However, the current technical possibilities seem to differ from what people want and expect from ATG. Future studies on the actual usage of ChatGPT will hopefully shed light on what people use it for indeed.

4.7. Limitations

Since this survey approach concentrated on depicting people's attitudes at two separate points in time, the two distinct samples do not allow for concluding inter-individual changes, thus rendering them as merely two snapshots. Furthermore, the second survey took place relatively shortly after the publication of ChatGPT. The German population might not have had enough time to get in touch with this tool, and different population groups did not have access to it equally quickly. As we are one of the first to systematically investigate people's attitudes and concepts regarding this specific subfield of AI, our survey captures rather general aspects. This also means that the aspects asked for, such as previous use of ChatGPT, were only recorded very superficially, meaning that large variance can be assumed at the individual level. Of course, the variety of technological implementations and possible applications could not be covered here by any means. Therefore, the cautious approach only allows for preliminary conclusions and should be specified in future studies.

4.8. Implications and Future Directions

The present study revealed that much of the German population has not yet had conscious contact with ATG technology, though the release of ChatGPT seemed to have an impact, at least on people's general awareness. On average, extreme attitudes were not observed. Whether this expresses a balanced attitude toward this specific AI application remains doubtful. The large proportion of answer selections expressing uncertainty in numerous scales and concepts instead shows that we confronted the samples with a relatively unknown field. It suggests a situation in which this type of AI flows into the most diverse areas at breakneck speed while facing a public widely naïve to it. Nevertheless, a certain amount of basic knowledge about ATG is present in some respondents, but the need for education also became apparent. Since our second survey took place only seven months after the release of ChatGPT and our results indicate that a significant part of the population still did not have the chance to get used to this technology, it remains to be observed how people's attitudes will change in the long run. Other analyses have made it clear that people with experience with this technology tend to have more positive attitudes and are more open to the use and consumption of ATG. Therefore, skepticism and unease should be encountered with broad knowledge and competence building.

The possible applications outside of journalism are as diverse as they are unexploited, ranging from creative writing tasks to highly formalized and standardized content generation and from informal interpersonal communication to academic writing. As ATG can be used in virtually any setting where text is required, future studies will have to cover a broad range of topics but also delve deep into specific subject matters. The diversity of technical approaches will lead to an improvement not only in ATG but also in general AI (cf. [57]). Given that the underlying data will be irrelevant for readers and that providers rarely label AI-generated text, public discussion should approach ATG with a different focus. Similarly,

for AI and algorithms in general, knowledge and competence building that reaches the general population [15] are necessary for ATG. Only with a sharpened understanding of chances and risks can users learn to handle ATG in an informed way and participate in a debate about potential regulations and the shaping of the technologies. The present study contributes to a basic understanding of current concepts and attitudes, which could serve as a benchmark for further studies. The results can also serve as a first indication for various types of stakeholders, who, for example, should orient on the desire for clear labeling and address misconceptions when letting people interact with AI-generated texts. Also, future research needs to understand what people expect from ATG and which misconceptions should be addressed. Speaking for the German population, with the current state of knowledge and awareness, the ground is paved for misattribution, disinformation, and credibility issues in journalism, while at the same time, building competence in informal and formal learning contexts cannot be fully exploited.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/bs14050353/s1, Figure S1: Relative frequencies (percentage) of participants' perceived probability on how likely they would be to read 18 topics written by AI, by study; Figure S2: Relative frequencies (percentage) of participants' author preference regarding 18 topics, by study; Table S1: Means and standard deviations of all variables separated by age group and study.

Author Contributions: Both A.L.H. and J.K. conceptualized the study, edited, and revised the manuscript. A.L.H. drafted the manuscript, conducted the studies, and monitored the recruitment of participants. A.L.H. analyzed and interpreted the data and is responsible for data curation. J.K. supervised the studies and was responsible for funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Leibniz-Institut für Wissensmedien (STB Data Science).

Institutional Review Board Statement: The study was conducted in accordance with the guidelines of the Local Ethics Committee of the Leibniz-Institut für Wissensmedien, which approved the study design and methods (Approval number: LEK 2023/022, approved on 17 May 2023).

Informed Consent Statement: Written informed consent was obtained from all participants involved in the study. Participants were invited to complete the online survey via the online market research platform Mingle in March 2022 (Study 1) and in June 2023 (Study 2).

Data Availability Statement: The data used in the study can be made available on requests addressed to the corresponding author.

Conflicts of Interest: The authors declare no competing interests.

References

1. Reiter, E.; Dale, R. Building applied natural language generation systems. *Nat. Lang. Eng.* **1997**, *3*, 57–87. [CrossRef]
2. Guzman, A.L. What is human-machine communication, anyway? In *Human-Machine Communication: Rethinking Communication, Technology, and Ourselves*; Guzman, A.L., Ed.; Peter Lang: New York, NY, USA, 2018; pp. 1–28.
3. McDonald, D.D. Issues in the choice of a source for natural language generation. *Comput. Linguist.* **1993**, *19*, 191–197.
4. Gatt, A.; Krahmer, E. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *J. Artif. Intell. Res.* **2018**, *61*, 65–170. [CrossRef]
5. Carlson, M. The robotic reporter: Automated journalism and the redefinition of labor, compositional forms, and journalistic authority. *Digit. J.* **2015**, *3*, 416–431. [CrossRef]
6. Clerwall, C. Enter the robot journalist: Users' perceptions of automated content. *J. Pract.* **2014**, *8*, 519–531. [CrossRef]
7. Köbis, N.; Mossink, L.D. Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Comput. Hum. Behav.* **2021**, *114*, 13. [CrossRef]
8. Lund, B.D.; Wang, T. Chatting about ChatGPT: How may AI and GPT impact academia and libraries? *Libr. Hi Tech News* **2023**, *40*, 26–29. [CrossRef]
9. Venkatesh, V.; Morris, M.G.; Davis, G.B.; Davis, F.D. User acceptance of information technology: Toward a unified view. *MIS Q.* **2003**, *27*, 425–478. [CrossRef]
10. Schick, N. *Deep Fakes and The Infocalypse: What You Urgently Need to Know*; Hachette UK: London, UK, 2020.

11. de Haan, Y.; van den Berg, E.; Goutier, N.; Kruikemeier, S.; Lecheler, S. Invisible friend or foe? How journalists use and perceive algorithmic-driven tools in their research process. *Digit. J.* **2022**, *10*, 1775–1793. [CrossRef]
12. Dörr, K.N. Mapping the Field of Algorithmic Journalism. *Digit. J.* **2016**, *4*, 700–722. [CrossRef]
13. Graefe, A. *Guide to Automated Journalism*; Tow Center for Digital Journalism: New York, NY, USA, 2016. [CrossRef]
14. Montal, T.; Reich, Z. I, Robot. You, Journalist. Who is the Author? Authorship, bylines and full disclosure in automated journalism. *Digit. J.* **2017**, *5*, 829–849. [CrossRef]
15. Overdiek, M.; Petersen, T. Was Deutschland über Algorithmen und Künstliche Intelligenz weiß und denkt. In *Ergebnisse Einer Repräsentativen Bevölverungsumfrage*; Bertelsmann Stiftung: Gütersloh, Germany, 2022.
16. Waddell, T.F. A robot wrote this? how perceived machine authorship affects news credibility. *Digit. J.* **2018**, *6*, 236–255. [CrossRef]
17. Graefe, A.; Haim, M.; Haarmann, B.; Brosius, H.-B. Readers' perception of computer-generated news: Credibility, expertise, and readability. *Journalism* **2018**, *19*, 595–610. [CrossRef]
18. Longoni, C.; Fradkin, A.; Cian, L.; Pennycook, G. News from artificial intelligence is believed less. *SSRN Electron. J.* **2021**. [CrossRef]
19. Van der Kaa, H.; Krahmer, E. Journalist versus news consumer: The perceived credibility of machine written wews. *Proc. Comput. J. Conf.* **2014**, *24*, 25.
20. Jia, C.; Johnson, T.J. Source credibility matters: Does automated journalism inspire selective exposure? *Int. J. Commun.* **2021**, *15*, 22.
21. Wölker, A.; Powell, T.E. Algorithms in the newsroom? News readers' perceived credibility and selection of automated journalism. *Journalism* **2021**, *22*, 86–103. [CrossRef]
22. Jang, W.; Chun, J.W.; Kim, S.; Kang, Y.W. The effects of anthropomorphism on how people evaluate algorithm-written news. *Digit. J.* **2021**. [CrossRef]
23. Tandoc, E.C.; Yao, L.J.; Wu, S. Man vs. machine? The impact of algorithm authorship on news credibility. *Digit. J.* **2020**, *8*, 548–562. [CrossRef]
24. Haim, M.; Graefe, A. Automated news: Better than expected? *Digit. J.* **2017**, *5*, 1044–1059. [CrossRef]
25. Wu, Y. Is automated journalistic writing less biased? An experimental test of auto-written and human-written news stories. *J. Pract.* **2020**, *14*, 1008–1028. [CrossRef]
26. Graefe, A.; Bohlken, N. Automated journalism: A meta-analysis of readers' perceptions of human-written in comparison to automated news. *Media Commun.* **2020**, *8*, 50–59. [CrossRef]
27. Sundar, S.S.; Kim, J. Machine heuristic: When we trust computers more than humans with our personal information. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, UK, 4–9 May 2019; pp. 1–9. [CrossRef]
28. Lermann Henestrosa, A.; Greving, H.; Kimmerle, J. Automated journalism: The effects of AI authorship and evaluative information on the perception of a science journalism article. *Comput. Hum. Behav.* **2023**, *138*, 107445. [CrossRef]
29. Longoni, C.; Cian, L. Artificial intelligence in utilitarian vs. hedonic contexts: The "word-of-machine" effect. *J. Mark.* **2022**, *86*, 91–108. [CrossRef]
30. Sundar, S.S. The MAIN model: A heuristic approach to understanding technology effects on credibility. In *Digital Media, Youth, and Credibility*; Metzger, J.M., Flanagin, J.A., Eds.; The MIT Press: Cambridge, MA, USA, 2008; pp. 73–100. [CrossRef]
31. Castelo, N.; Bos, M.W.; Lehmann, D.R. Task-dependent algorithm aversion. *J. Mark. Res.* **2019**, *56*, 809–825. [CrossRef]
32. Bigman, Y.E.; Gray, K. People are averse to machines making moral decisions. *Cognition* **2018**, *181*, 21–34. [CrossRef] [PubMed]
33. Böhm, R.; Jörling, M.; Reiter, L.; Fuchs, C. Content beats competence: People devalue ChatGPT's perceived competence but not its recommendations. *PsyArXiv* **2023**. [CrossRef]
34. nextMedia.Hamburg. 2018. Available online: https://www.nextmedia-hamburg.de/wp-content/uploads/2019/02/20180809_journalismusderzukunft.pdf (accessed on 1 February 2024).
35. nextMedia.Hamburg. 2019. Available online: https://www.nextmedia-hamburg.de/wp-content/uploads/2019/08/nextMedia-Umfrage_KI_2019_PM-1.pdf (accessed on 1 February 2024).
36. Kennedy, B.; Tyson, A.; Saks, E. Public Awareness of Artificial Intelligence in Everyday Activities. 2023. Available online: https://policycommons.net/artifacts/3450412/public-awareness-of-artificial-intelligence-in-everyday-activities/4250673/ (accessed on 1 February 2024).
37. Fischer, S.; Petersen, T. Was Deutschland über Algorithmen weiß und Denkt: Ergebnisse einer Repräsentativen Bevölkerungsumfrage. 2018. Available online: https://www.bertelsmann-stiftung.de/de/publikationen/publikation/did/was-deutschland-ueber-algorithmen-weiss-und-denkt (accessed on 1 February 2024).
38. Venkatesh, V.; Davis, F.D. A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Manag. Sci.* **2000**, *46*, 186–204. [CrossRef]
39. Darda, K.; Carre, M.; Cross, E. Value attributed to text-based archives generated by artificial intelligence. *R. Soc. Open Sci.* **2023**, *10*, 220915. [CrossRef] [PubMed]
40. Zhang, B.; Dafoe, A. Artificial Intelligence: American Attitudes and Trends. 2019. Available online: https://isps.yale.edu/sites/default/files/files/Zhang_us_public_opinion_report_jan_2019.pdf (accessed on 1 February 2024).
41. Mays, K.K.; Lei, Y.; Giovanetti, R.; Katz, J.E. AI as a boss? A national US survey of predispositions governing comfort with expanded AI roles in society. *AI Soc.* **2021**. [CrossRef]

42. Cave, S.; Coughlan, K.; Dihal, K. "Scary Robots" Examining Public Responses to AI. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Honolulu, HI, USA, 27–28 January 2019; pp. 331–337. [CrossRef]
43. Kimmerle, J.; Timm, J.; Festl-Wietek, T.; Cress, U.; Herrmann-Werner, A. Medical Students' Attitudes toward AI in Medicine and their Expectations for Medical Education. *J. Med. Educ. Curric. Dev.* **2023**, *10*, 23821205231219346. [CrossRef] [PubMed]
44. Ada Lovelance Institute & Alan Turing Institute. How Do People Feel about AI? A Nationally Representative Survey of Public Attitudes to Artificial Intelligence in Britain. 2023. Available online: https://adalovelaceinstitute.org/report/public-attitudes-ai (accessed on 1 February 2024).
45. Schepman, A.; Rodway, P. Initial validation of the general attitudes towards Artificial Intelligence Scale. *Comput. Hum. Behav. Rep.* **2020**, *1*, 100014. [CrossRef] [PubMed]
46. Said, N.; Potinteu, A.E.; Brich, I.R.; Buder, J.; Schumm, H.; Huff, M. An artificial intelligence perspective: How knowledge and confidence shape risk and opportunity perception. *Comput. Hum. Behav.* **2022**. [CrossRef]
47. Géron, A. *Hands-On Machine Learning Aith Scikit-Learn, Keras, and TensorFlow*; O'Reilly Media: Sebastopol, CA, USA, 2022.
48. Arlot, S.; Celisse, A. A survey of cross-validation procedures for model selection. *Stat. Surv.* **2010**, *4*, 40–79. [CrossRef]
49. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*; Springer: Berlin/Heidelberg, Germany, 2013; Volume 112.
50. Wu, J.; Du, H. Toward a better understanding of behavioral intention and system usage constructs. *Eur. J. Inf. Syst.* **2012**, *21*, 680–698. [CrossRef]
51. Giering, O.; Fedorets, A.; Adriaans, J.; Kirchner, S. *Künstliche Intelligenz in Deutschland: Erwerbstätige Wissen oft Nicht, dass sie mit KI-Basierten Systemen Arbeiten*; Deutsches Institut für Wirtschaftsforschung e.V.: Berlin, Germany, 2021; Volume 88, pp. 783–789.
52. Abdaljaleel, M.; Barakat, M.; Alsanafi, M.; Salim, N.A.; Abazid, H.; Malaeb, D.; Mohammed, A.H.; Hassan, B.A.R.; Wayyes, A.M.; Farhan, S.S. A multinational study on the factors influencing university students' attitudes and usage of ChatGPT. *Sci. Rep.* **2024**, *14*, 1983. [CrossRef] [PubMed]
53. Broos, A. Gender and information and communication technologies (ICT) anxiety: Male self-assurance and female hesitation. *CyberPsychology Behav.* **2005**, *8*, 21–31. [CrossRef] [PubMed]
54. Bodani, N.; Lal, A.; Maqsood, A.; Altamash, S.; Ahmed, N.; Heboyan, A. Knowledge, Attitude, and Practices of General Population Toward Utilizing ChatGPT: A Cross-sectional Study. *SAGE Open* **2023**, *13*, 21582440231211079. [CrossRef]
55. Proksch, S.; Schühle, J.; Streeb, E.; Weymann, F.; Luther, T.; Kimmerle, J. The Impact of Text Topic and Assumed Human vs. AI Authorship on Competence and Quality Assessment. 2024. Available online: https://osf.io/preprints/osf/7fhwz (accessed on 1 February 2024).
56. Lermann Henestrosa, A.; Kimmerle, J. The Effects of Assumed AI vs. Human Authorship on the Perception of a GPT-Generated Text. 2024. Available online: https://osf.io/preprints/psyarxiv/wrusc (accessed on 1 February 2024).
57. Sun, X.; Zhang, J.; Wu, X.; Cheng, H.; Xiong, Y.; Li, J. Graph prompt learning: A comprehensive survey and beyond. *arXiv* **2023**, arXiv:2311.16534.

B.4 Manuscript IV

Lermann Henestrosa, A., & Kimmerle, J. (2024). Data Descriptor for "Understanding and Perception of Automated Text Generation among the Public: Two Surveys with Representative Samples in Germany". *Data, 9*(10), Article 116. https://doi.org/10.3390/data9100116

data

MDPI

Data Descriptor

Data Descriptor for "Understanding and Perception of Automated Text Generation among the Public: Two Surveys with Representative Samples in Germany"

Angelica Lermann Henestrosa [1,*] and Joachim Kimmerle [1,2]

[1] Knowledge Construction Lab, Leibniz-Institut für Wissensmedien, 72076 Tübingen, Germany; j.kimmerle@iwm-tuebingen.de
[2] Department of Psychology, Eberhard Karls University, 72076 Tübingen, Germany
* Correspondence: a.lermann-henestrosa@iwm-tuebingen.de

Abstract: With the release of ChatGPT, text-generating AI became accessible to the general public virtually overnight, and automated text generation (ATG) became the focus of public debate. Previously, however, little attention had been paid to this area of AI, resulting in a gap in the research on people's attitudes and perceptions of this technology. Therefore, two representative surveys among the German population were conducted before (March 2022) and after (July 2023) the release of ChatGPT to investigate people's attitudes, concepts, and knowledge on ATG in detail. This data descriptor depicts the structure of the two datasets, the measures collected, and potential analysis approaches beyond the existing research paper. Other researchers are encouraged to take up these data sets and explore them further as suggested or as they deem appropriate.

Dataset: https://osf.io/sn75h/.

Keywords: automated text generation; public attitudes toward AI; ChatGPT impact; automated journalism; conceptions on AI

Citation: Lermann Henestrosa, A.; Kimmerle, J. Data Descriptor for "Understanding and Perception of Automated Text Generation among the Public: Two Surveys with Representative Samples in Germany". *Data* **2024**, *9*, 116. https://doi.org/10.3390/data9100116

Academic Editor: Han Woo Park

Received: 11 September 2024
Revised: 3 October 2024
Accepted: 9 October 2024
Published: 11 October 2024

1. Summary

Recent developments in the availability of automated text generation (ATG) technology have a major impact on many areas of society, ranging from education to science, economics, journalism, and media. The adequate usage and adoption of LLMs (Large language models) like GPT, Claude, or Llama can depend on what people believe their capabilities are. Moreover, people's attitudes influence the acceptance of AI for text generation, which has been integrated into many areas. It is, therefore, of great importance to understand what people know about these developments and how they perceive and evaluate them. The data descriptor presented here comprises two sets of survey data collected at two distinct time points (March 2022 and July 2023). The data were collected from two different German samples to capture the current attitudes, concepts, and experiences related to the technology at each of these moments. Therefore, the first survey depicts those measures before the release of ChatGPT, while the second survey was conducted after its launch, which happened in November 2022. The second survey is an exact replication of the first, with the addition of variables to account for people's experiences with this new technology, i.e., a knowledge test, experience with ChatGPT, attitudes toward ChatGPT, and lay attitudes toward ATG and ChatGPT.

These data sets mark two pivotal moments in capturing specific attitudes and concepts toward ATG, i.e., before the public focus on large language models (LLM) and shortly after the release of the first open access LLM-based Chatbot. As such, they provide a

valuable benchmark for future studies on people's attitudes and perceptions, offering two snapshots at the earliest possible points in time. Moreover, as the surveys were carried out on large representative samples, we invite researchers to analyze these data sets in more detail and to look at potential subgroups, which we have not conducted in the connected publication [1].

Both surveys were distributed online via Mingle (Bilendi, Cologne, Germany) to collect data from German samples, which were representative of age, gender, and educational level (see Table 1).

Table 1. Composition of the final samples regarding gender, age, and educational level.

Quota		Survey 1		Survey 2	
		Target Quota	Final Composition	Target Quota	Final Composition
gender	Female	49.82	51.26	49.54	50.94
	male	50.18	48.54	50.46	48.86
	diverse	0.91	0.19	-	0.20
age	18–29 year	18.91	17.90	19.76	18.56
	30–39 year	18.00	17.70	19.40	18.56
	40–49 year	16.91	17.02	17.97	18.56
	50–59 year	22.36	22.57	23.74	24.48
	60–69 year	23.64	24.81	19.12	19.84
education *	low	34.00	33.66	27.00	26.65
	middle	32.00	31.52	33.00	34.45
	high	34.00	34.82	41.00	38.89

* Educational levels were built as follows: no or elementary/primary school degree = low, secondary school degree = middle, high school degree = high.

This project was funded by the Leibniz-Institut für Wissenmedien, STB Data Science, in Germany and is part of the project "AI for science communication: Acceptance and laypeople comprehension". The main results are published in the journal *Behavioral Sciences* (https://doi.org/10.3390/bs14050353, accessed on 03 October 2024).

2. Data Description

On the platform of the Open Science Framework (OSF), two separate primary data files in RDS format are stored for each survey. Moreover, an analysis R file is stored, in which these two data files are merged as one of the first steps. However, they can be analyzed separately, especially since the second sample consists of different participants.

In the data files, each row is a participant, and each column is a variable. For the variable names and their corresponding column names in the data sets, see Table 2. As most scales consisted of multiple items that are not yet aggregated or further processed in the data set for reasons of transparency, the columns in the data set are provided with the suffix "_[Item number]" and represent the single items of one measure. The codebook attached to the OSF project contains a detailed description of all variables, as well as all items and their choice options.

We recommend analyzing the data using R or R-studio. However, it is also possible to import the data set to other statistical analysis programs. To analyze the data with our provided script, it is necessary to install R or R-Studio in version 4.3.1 (16 June 2023 ucrt) or newer. The necessary packages are listed in the beginning of the analysis script and need to be installed in advance.

Table 2. Overview of all variables used in the surveys.

Variable	Number of Items	Scale Formation	Cronbach Alpha Survey 1/2	Scale Type	Example Item/Choice Option	Answer Options	Source	Column Name
Self-assessed knowledge	1	-	-	Ordinal	How would you rate your level of knowledge regarding AI?	No—Limited—Some—Good—Extensive	Self-developed	SELFSKILL
General attitudes °	20	mean	0.91/0.92	Likert	AI is exciting.	Strongly disagree—Rather disagree—neutral—Rather agree—Strongly agree	Adopted from [2]	GENATT
Machine heuristic	4		0.75/0.79	Likert	When an AI performs a task, it is—accurate.	Strongly disagree—Rather disagree—Partly/partly—Rather agree—Strongly agree	Adopted from [3]	MACHEU
Media use [1]	4/5	-	-	Ordinal	Which of the following voice-based applications do you use and how often do you use them?—chatbots	Never—Seldom—Occasionally—Often—Constant	Self-developed	USE_1:_6
Media use for scientific information retrieval [2]	11/14	-	-	Ordinal	How often do you use the following media to obtain information on scientific topics?—blogs		Self-developed	MEDIA_USE_1:_14
Heard about ATG	1	-	-	Ordinal	Have you ever heard that AIs can write texts?		Self-developed	HEARD
Read AI text	1	-	-	Ordinal	Have you ever read texts written by an AI?		Self-developed	READ
Type of AI text read	1	-	-	Text entry	What kind(s) of text(s) have you already read? Please describe them briefly.		-	AITXT

Table 2. *Cont.*

Variable	Number of Items	Scale Formation	Cronbach Alpha Survey 1/2	Scale Type	Example Item/Choice Option	Answer Options	Source	Column Name
Knowledge *	15	Sum score of correctly answered items	-	Nominal	AI language models (e.g., ChatGPT) calculate for their answers which word is most likely to come next.	Right—Wrong—Don't know	Partly self-developed, partly adapted from [4]	KNOW_1:_15
Function of ATG †	5/6	proportions	-	Likert	The AI uses existing words and texts and reassembles them.	Not at all—Rather unlikely—Perhaps—Rather likely—For certain	Self-developed	FUNC_1:_6
Sources of ATG †	6/7		-	Likert	The AI generates the content itself, without human intervention.			SOURC_1:_7
Human control †	4/5		-	Likert	The human sees the end product and edits the text if necessary.			CONTR_1:_5
Responsible	9		-	Likert	The programmer			RESP_1:_9
Performance expectancy	3	mean	0.78/0.81	Likert	AI-written texts would make me more productive.	Strongly disagree—Rather disagree—Partly/partly—Rather agree—Strongly agree	Adapted from [5]	TAM_1:_3
Effort expectancy	3		0.71/0.71	Likert	I think AI-written texts are clear and understandable.			TAM_4:_6
Attitude toward using the technology °	4		0.78/0.79	Likert	I would like to read AI-generated texts.			TAM_7:_10
Anxiety	3		0.71/0.68	Likert	I have reservations about reading AI-written texts.			TAM_11:_13
Behavioral intention °	3		0.64/0.67	Likert	I intend to read AI-written texts in the future.			TAM_14:_16
Permission to write like human °	4		0.63/0.62	Likert	AIs should be allowed to write about the same topics as humans.			TAM_17:_20
Intentions to read AI texts	18	proportions	-	Likert	Please indicate how likely you would be to read AI-written texts on the following topics.—Politics	Not at all—Rather unlikely—Perhaps—Rather likely—For certain	Self-developed	INTEN_1:_18
Comparison human vs. AI	18	proportions	-	Likert	If you had the choice, who would you prefer to be informed about the following topics?—Politics	Rather from a human—No preference—Rather from an AI		COMP_1:_18

Table 2. *Cont.*

Variable	Number of Items	Scale Formation	Cronbach Alpha Survey 1/2	Scale Type	Example Item/Choice Option	Answer Options	Source	Column Name
ChatGPT use *	1	-	-	Text entry	You stated at the beginning that you have used ChatGPT before. Please describe what you have used or are using ChatGPT for.	-	Self-developed	USEGPT
Attitudes toward ChatGPT *,°	16	mean	-/0.89	Likert	I am satisfied with the answers from ChatGPT.	Strongly disagree—Rather disagree—Partly/partly—Rather agree—Strongly agree	Self-developed	ATTGPT_1:_16
Lay attitudes toward ATG and ChatGPT *,°	9		-/0.82	Likert	I am optimistic about the impact of automated text generation (e.g., ChatGPT) on society.		Self-developed	LAYGPT_1:_9
Pro and contra arguments	1	-	-	Text entry	You now have the opportunity to freely express your thoughts on the topic.For example: What advantages and disadvantages do you see in automated text creation using AI?	-	Self-developed	PROCON

* Only measured in Survey 2. ° Contain reverse coded items, see R script. [1] In Survey 2, the option "ChatGPT" was added. [2] In Survey 2, the options "ChatGPT", "Books", and "Podcasts" were added. † In Survey 2, an additional item to those scales was added, which captured the technology behind ChatGPT.

3. Methods

3.1. Survey Design

The surveys were developed to capture people's experience, attitudes, concepts, knowledge, and preferences toward ATG and ChatGPT, and to examine whether there were changes between these two time points.

In each survey, all participants were presented with the full set of questions, with one exception: in the beginning, participants were asked about their experience with ATG. If they indicated that they had previously read an AI-written text, they were directed to an open text entry question, where they could specify the type(s) of text they had read. This question created missing values for a proportion of respondents. The same procedure was carried out in Survey 2. However, in the second survey, we had the chance to ask specifically about their experience with ChatGPT, which again generated a subset of participants with ChatGPT experience who were later forwarded to the attitudes toward ChatGPT scale. Therefore, individuals who indicated no experience with ChatGPT have missing values at this measure. For each scale, the items were presented in random order to avoid order effects.

3.2. Survey Platform

The data collection and questionnaire setup were conducted using Qualtrics (Qualtrics LLC; CEO: Ryan Smith; Address: 2250 N. University Pkwy, 48-C, Provo, UT 84604, USA). Qualtrics was chosen for its robust features, including customizable survey design and secure data storage.

3.3. Participant Recruitment

The panel provider *Bilendi and respondi* was commissioned with the recruitment of the samples. It is ISO certified (ISO 20252:2019) and processes personal data in accordance with the European General Data Protection Regulation (DSGVO). It sent out invitations via its panel Mingle. The provider was also responsible for quota management. For the quotas, according to which participants were recruited, and the composition of the final samples, see Table 1. According to Mingle, participant compensation follows a points system in which points can be collected depending on the length and type of the study. The participants themselves decide whether to convert the points into cash, vouchers, or donations. For participating in Survey 1, respondents received the equivalent of EUR 1.30 for a mean duration of 11.77 min. In Survey 2, they received EUR 1.93 for a mean duration of 15.52 min.[1]

3.4. Data Filtering

In Survey 1, from 1111 participants who completed the questionnaire, 83 people did not give their consent to data use. Thus, the final sample size resulted in N = 1028. In Survey 2, from 1116 participants who filled in the questionnaire completely, 51 withdrew their data at the end of the survey. Another 101 persons were excluded due to an incorrect answer to the attention check implemented in the knowledge test (item "KNOW_16"). Therefore, the final sample in Survey 2 consisted of N = 1013 respondents.

3.5. Data Curation and Storage

The original data from both surveys are archived on the local servers of the Leibniz-Institut für Wissensmedien for at least ten years after publication. The prepared and fully anonymized primary data are publicly available at OSF. The data stored at Qualtrics will be deleted by 31 March 2025, at the latest. Qualtrics complies with the European guidelines on data storage. Bilendi and respondi had no access to the survey data during data collection and can therefore not connect the participant data to the survey data. In turn, we as authors had no access to Bilendi and respondi's participant data at any point in time.

There are no missing values in both data sets as the data were only analyzed if participants fully completed the surveys and gave their written informed consent to using

their data for research purposes on the last page. Furthermore, participants in Survey 2 were excluded from the analysis if they incorrectly answered the attention check item implemented in the knowledge test.

3.6. Measures and Procedure

Table 2 shows all measures used in the surveys in the order they were presented. The original Qualtrics questionnaire can be found under the material here: https://osf.io/sn75h/ (accessed on 03 October 2024). The original data set also contains variables for the exact age, education, and occupation. Note that participants entered their occupations in an open text entry, which is why the variable requires further processing. Please see [1] or the R script for the correct answers to the knowledge test.

4. Limitations

The present data set contains two snapshots of different samples at two points in time and is, therefore, unsuitable for analyzing intrapersonal changes. Since the surveys were conducted with samples from the German population, any comparisons or inferences regarding other countries should be approached with caution. Consequently, the scale translations need to be validated when replicating these studies in different languages.

Moreover, since the first survey was conducted before the public awareness caused by the launch of ChatGPT, the concepts and scales can still be interpreted as too general. The concepts of functionality, data retrieval, responsibility, and human control are not tailored to specific technologies but capture different aspects that apply more or less to various systems. Furthermore, there is certainly high variability in the answer options "perhaps" of the concepts and the answer option "don't know" of the knowledge test, for which this data set cannot provide explanations. For example, future research could focus on the most frequently used AI systems and adapt the questions accordingly. Open answer options could offer further insights into the sources of uncertainties or gaps in knowledge.

5. Potential Applications of the Data Set

This open-access data set can be seen as an important starting point for subsequent research. This data set not only provided a baseline measure of attitudes toward ATG before public attention was drawn to this technology, but also offered a broad overview of various constructs that may become relevant when ATG and LLMs gain wider adoption in the general population. Future research may also delve deeper into specific misconceptions, risks, and opportunities that the population perceives. The following research questions might serve scientists in this research field in using this data set for exploratory purposes. Potential areas of further research include socio-demographic aspects, which can provide further information about differences between population groups and thus enrich our knowledge about a potential AI divide:

RQ1: Do attitudes, preferences, or knowledge differ regarding age, gender, or educational background?

Moreover, further investigation of the relationship between personal attitudes and behavioral intentions can provide information about factors that determine actual usage:

RQ2: How well do lay attitudes toward ChatGPT, attitudes toward using the technology, and performance expectancy predict behavioral intentions?
RQ3: How well do the TAM/UTAUT subscales predict behavioral intention and ChatGPT use in Survey 2?

To our knowledge, no scale captures specific attitudes toward ATG technology or measures adequate to differentiate between users and non-users or even different levels of experience. Therefore, our self-developed measurements can serve as a basis for scale development and validation:

RQ4: How valid is the newly developed scale in measuring lay attitudes toward ChatGPT? Can it be adapted to LLMs in general?

RQ5: How valid is the newly developed scale in measuring attitudes toward ChatGPT?
RQ6: Do participants with and without ChatGPT experience differ in their ratings?
RQ7: What do the open answers on ChatGPT usage tell us?

Author Contributions: Both A.L.H. and J.K. conceptualized this study, edited, and revised the manuscript. A.L.H. drafted the manuscript, conducted the studies, and monitored the recruitment of participants. A.L.H. analyzed and interpreted the data and is responsible for data curation. J.K. supervised the studies and was responsible for funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Leibniz-Institut für Wissensmedien, STB Data Science.

Institutional Review Board Statement: This study was conducted in accordance with the guidelines of the Local Ethics Committee of the Leibniz-Institut für Wissensmedien, which approved the study design and methods (Approval number: LEK 2023/022, approved on 17 May 2023).

Informed Consent Statement: Written informed consent was obtained from all participants involved in this study. Participants were invited to complete the online surveys via the online market research platform Mingle in March 2022 (Study 1) and in June 2023 (Study 2).

Data Availability Statement: The data, analysis script, and material are openly available in OSF at https://osf.io/sn75h/ (accessed on 03 October 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

Note

[1] Mean duration was calculated by excluding outliers detected via the interquartile range (IQR) method. Following this, we defined outliers as observations that fell below Q1 (first quartile) − 1.5 × IQR or above Q3 (third quartile) + 1.5 × IQR. Thus, for Survey 1, the data from 75 respondents whose duration was equal to or longer than 28.27 min were excluded from the calculation of the average duration. For Survey 2, the data from 98 participants whose duration was equal to or longer than 40.52 min were not included in the average duration calculation. Note that extreme outliers regarding duration of survey completion can be caused by missing the submission of the last survey page or breaks during the survey.

References

1. Lermann Henestrosa, A.; Kimmerle, J. Understanding and Perception of Automated Text Generation among the Public: Two Surveys with Representative Samples in Germany. *Behav. Sci.* **2024**, *14*, 353. [CrossRef] [PubMed]
2. Schepman, A.; Rodway, P. Initial Validation of the General Attitudes towards Artificial Intelligence Scale. *Comput. Hum. Behav. Rep.* **2020**, *1*, 100014. [CrossRef] [PubMed]
3. Sundar, S.S.; Kim, J. Machine Heuristic: When We Trust Computers More than Humans with Our Personal Information. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, Scotland, 4–9 May 2019; pp. 1–9. [CrossRef]
4. Said, N.; Potinteu, A.E.; Brich, I.R.; Buder, J.; Schumm, H.; Huff, M. An Artificial Intelligence Perspective: How Knowledge and Confidence Shape Risk and Opportunity Perception. *Comput. Hum Behav.* **2022**, *149*, 107855. [CrossRef]
5. Venkatesh, V.; Morris, M.G.; Davis, G.B.; Davis, F.D. User Acceptance of Information Technology: Toward a Unified View. *MIS Q.* **2003**, *27*, 425–478. [CrossRef]

B.5 Manuscript V

Lermann Henestrosa, A., & Kimmerle, J. (2025). "Always Check Important Information!" The Role of Disclaimers in the Perception of AI-generated Content. *Computers in Human Behavior: Artificial Humans, 4*, Article 100142. https://doi.org/10.1016/j.chbah.2025.100142

Computers in Human Behavior: Artificial Humans 4 (2025) 100142

Contents lists available at ScienceDirect

Computers in Human Behavior: Artificial Humans

journal homepage: www.journals.elsevier.com/computers-in-human-behavior-artificial-humans

"Always check important information!" - The role of disclaimers in the perception of AI-generated content

Angelica Lermann Henestrosa [a,*], Joachim Kimmerle [a,b]

[a] Knowledge Construction Lab, Leibniz-Institut Fuer Wissensmedien, Tübingen, Germany
[b] Department of Psychology, Eberhard Karls University, Tübingen, Germany

ARTICLE INFO

Keywords:
Disclaimer
System transparency
Explainable AI (XAI)
Generative artificial intelligence
Credibility perceptions
Science communication

ABSTRACT

Generative AI, and large language models (LLMs) in particular, have become a prevalent source of digital content. Despite their widespread availability, these models come with critical weaknesses, such as a lack of factual accuracy. Being informed about the advantages and disadvantages of these tools is essential for using AI safely and adequately, yet not everyone is aware of them. Therefore, we explored in three experimental studies how disclaimers affect people's perceptions of AI-authorship and AI-generated content on scientific topics. Additionally, we investigated the impact of information presentation and authorship attributions—whether content is authored solely by AI or co-authored with humans. Across the experiments, no effects of disclaimer type on text perceptions and only minor effects on authorship perceptions were found. In Study 1, an evaluative (vs. neutral) information presentation decreased credibility perceptions, while informing about AI's strengths vs. limitations did not. In addition, we found participants to believe in the machine heuristic, that is, to attribute more accuracy and less bias to AI than to human authors. Study 2 revealed interaction effects between authorship and disclaimer type, providing insights into possible balancing effects of human-AI co-authorship. In Study 3, both strengths and limitations disclaimers induced higher credibility ratings than basic disclaimers. This research suggests that disclaimers fail to univocally influence the perception of AI-generated output. Further interventions should be developed to raise awareness of the capabilities and limitations of LLMs and to advocate for ethical practices in handling AI-generated content, especially regarding factual information.

1. Introduction

Generative artificial intelligence (GenAI) refers very broadly to "models that may produce new, previously unseen information dependent on the data on which they were trained" (García-Peñalvo & Vázquez-Ingelmo, 2023, p. 7) and has spread enormously, especially since the release of Large Language Models (LLMs) like ChatGPT. As GenAI seems to be here to stay, it will likely be seamlessly integrated into nearly every device people communicate with, leading to more natural, subtle, and unnoticed interactions with GenAI. People already use text and image GenAI intensively for a lot of different purposes: for scientific information search (Greussing et al., 2024; Lermann Henestrosa & Kimmerle, 2024), answering factual questions (Fletcher & Nielsen, 2024), self-diagnosis (Shahsavar & Choudhury, 2023), or even research (Hosseini et al., 2023). However, although LLMs like Llama, Claude, or ChatGPT display unprecedented linguistic capabilities, they have significant disadvantages, for example, regarding factual accuracy (Wei et al., 2024).

Even though the factual accuracy rates tend to improve with every new model, they will hardly reach 100 % since the inherent unpredictability and the amount of data that makes the linguistic feats possible in the first place excludes factual reviews of every piece of information. Therefore, no matter how strongly post-hoc adjustments can reduce the error rate, there are areas like new research fields, in which the provision of incorrect information by LLMs is very likely and thus highly problematic. Moreover, LLMs express themselves in a very self-confident way. This is especially relevant in the case of critical and sensitive material such as health information, for instance, where factual accuracy must be the benchmark and for which studies show that the error rate is considerably high, especially on less established scientific topics (Bulian et al., 2023; Deiana et al., 2023; Wei et al., 2024). Inconsistencies, minor errors, or incorrect citations in long continuous texts that would require subsequent research for verification of the provided information can hardly be identified as such by laypersons in

* Corresponding author. Leibniz-Institut fuer Wissensmedien, Schleichstraße 6, D-72076, Tübingen, Germany.
E-mail address: a.lermann-henestrosa@iwm-tuebingen.de (A. Lermann Henestrosa).

https://doi.org/10.1016/j.chbah.2025.100142
Received 29 December 2024; Received in revised form 11 March 2025; Accepted 19 March 2025
Available online 22 March 2025

the respective area. Therefore, users are predestined to take incomplete information, or even misinformation, at face value. Moreover, the eloquent and persuasive presentation of such information could encourage people to use that information uncritically and ascribe capabilities to LLMs that do not match their actual functionalities.

Indeed, recent studies have shown that people hold different misconceptions regarding the capabilities of AI and lack knowledge about how LLMs function (Hornberger et al., 2023; Lermann Henestrosa & Kimmerle, 2024; Said et al., 2022). While, for example, 56 % of participants in a survey among the German population knew that the statements of language-generating AI can be incorrect, 47 % believed it to be true that LLMs have learned to understand language like a human (Lermann Henestrosa and Kimmerle, 2024). Moreover, this study found a general tendency to ascribe the adjectives "accurate", "unbiased", "neutral", and "objective" to an AI, an assessment also known as *machine heuristic* (Sundar, 2008, pp. 1–9). However, in the case of LLMs that are consulted on scientific topics, the attribution of such characteristics could also represent an overestimation of their actual abilities.

Consequently, if LLMs are used for tasks for which they were not technically designed, this can lead to the replication and dissemination of misinformation. Thus, it is important to educate users about how anLLM's answers are generated and how reliable they are. One option is to provide people with specific information about the strengths and limitations of the technology they're about to interact with. This could be done by presenting *disclaimers* before the consumption of articles generated by AI. Disclaimers are snippets of information intended to draw people's attention to the dangers or risks of a system and can be used by website operators to exclude liability. Currently, the companies behind LLMs mention limitations, at least in bylines on their interfaces. For instance, OpenAI uses a disclaimer appearing at the ChatGPT interface as a small footnote stating: "ChatGPT can make mistakes. Always check important information" (as of March 2025). Besides the legal matter of safeguarding against potentially false information, the question arises of whether users even consider these warnings appropriately. If so, how they process them and if this influences their perception of generated output in the case of LLMs remains largely unexplored. Going a step further, it is crucial to examine not only if and what information is provided but also which specific details enable users best to engage with LLMs responsibly. While emphasizing the essential facts that users must be aware of, it must also be kept reasonably short for effective comprehension. A more detailed explanation of LLMs' capabilities and limitations wherever it is being applied is an obvious intervention but one that lacks validation. Beyond legal considerations, informing users about these aspects could also help to design optimized and safer interaction with these systems. This is crucial, given the wide range of possible uses and applications, where "checking important information" may not always be intuitive, feasible, or conceivable.

An open question, which comes before the question of exactly what information must be presented in what way, is whether prior information influences perception at all. Therefore, the research presented here is dedicated to examining different disclaimer types on the evaluation of AI-generated content and the AI itself. Furthermore, since purely AI-generated content plays a role mainly in direct interaction with GenAI, but collaboration between humans and AI is also increasingly conceivable in other contexts when writing texts, we also investigated if co-authorship affects those perceptions compared to pure AI authorship. A context in which a human verification authority is particularly important is science communication. Here, the perception of different AI author variations and the awareness on information to be critically scrutinized is particularly important to research.

1.1. Research on the effects of disclaimers

The effects of information given about a system have been investigated under multiple different terms, like system transparency (Wang & Huang, 2024), algorithmic transparency (Rader et al., 2018, pp. 1–13), explainable AI (Xu et al., 2019), or AI framing (Seo et al., 2019, pp. 265–274). The explainable AI (XAI) approach initially aimed to provide people with information on how AI arrived at its decisions and thus help people make more informed decisions when using such an algorithm (Alufaisan et al., 2021). However, research has shown that such explanations do not always lead to better decisions (e.g., Bansal et al., 2021, pp. 1–16; Jacobs et al., 2021) and can potentially harm trust (Papenmeier et al., 2019). In addition, the case of presentation of scientific information, for instance, is not fully comparable with algorithmic decision-making since an action by the user does not necessarily follow the consumption of factual information. The uptake of information is thus not linked to a deliberate decision-making process and is inherently more passive.

In automated journalism, transparency is broadly defined as "the disclosure of information about algorithms to enable monitoring, checking, criticism, or intervention by interested parties" (Diakopoulos & Koliska, 2017, p. 811). While this could be a maxim for AI-supported journalism, transparency communicated via a disclaimer could be less practicable for lay users of GenAI. In general, a disclaimer is a declaration in which a company distances itself from specific contents for legal reasons. In this paper, we specifically refer to its informative purpose, which is to draw attention to certain characteristics of AI. Providing users with context and explanations about a technology can alleviate uncertainty, enhance trust (Liu, 2021), help understand how AI works, and how to assess if output is biased (Phillips et al., 2021; Rader et al., 2018, pp. 1–13). Furthermore, Liu (2021) argues that transparency itself might function as a heuristic cue, leading to a more positive perception of the AI independent of how deeply the explanation's quality might be examined (Shin & Park, 2019). In connection to this, Lee (2024) raises the question of whether interaction with AI is fundamentally mindless, as this assumption is repeatedly used in the CASA paradigm (Nass et al., 1994; Nass & Moon, 2000) to explain people's social and thoughtless reactions to AI.

However, there are mixed findings on the effectiveness of prior information. In a recent meta-analysis, Wang and Huang (2024) found that in the presence of system transparency (i.e., if participants were informed that an algorithm had produced the news and how it functioned), AI-written news induced lower credibility perceptions than human-authored ones. On the contrary, findings from the field of explainable AI suggest that explaining the AI might increase trust in it (Shin, 2021; Vered et al., 2023). However, a study by Metzger et al. (2024), who compared different disclaimer types, could not find any effect of prior information on trust perceptions. These conflicting results highlight a critical research gap: While attempts at transparency are used and can be a valuable contribution to the safe and responsible use of AI, their actual impact on user perceptions remains inconclusive. Furthermore, research on the comparison of different disclaimer types and their influence on perceived credibility of the presented content is scarce. In addition, it is still unclear how disclaimers influence other aspects of the perception of AI. Such could be attributed intelligence and anthropomorphism, as information about the weaknesses of a tool could impact how intelligent people perceive the AI to be. Also, the question arises to what extent users distinguish between human and AI authors regarding perceived anthropomorphism. Anthropomorphism is defined as "the tendency to imbue the real or imagined behavior of nonhuman agents with humanlike characteristics, motivations, intentions, or emotions" (Epley et al., 2007, p. 864), a concept highly investigated in chatbot research. In the context of science communication, previous research has found that a human author is perceived as more anthropomorphic compared to an AI author, although this does not consistently translate to being perceived as more intelligent (Lermann Henestrosa et al., 2023). Since current LLMs represent a sophisticated evolution of conventional chatbots, particularly in terms of anthropomorphic writing style, it is important to consider how information about a lack of genuine understanding, for instance, might influence this perception.

Therefore, the objective of this work is to investigate whether and how informative disclaimers about AI capabilities influence the perceptions of AI output and AI as an author in science communication. Moreover, we aim to examine the effects of information presentation and co-authorship in connection with different disclaimer types. Specifically, we sought to answer the following research questions.

- How do disclaimers that emphasize AI's capabilities (e.g., the strengths or limitations) influence text perceptions, including message credibility and behavioral intentions to consume such content?
- How does information presentation (neutral vs. evaluative) affect perceptions of the content and its author? Does information presentation interact with disclaimer type?
- How do disclaimers shape perceptions of the author, specifically in terms of perceived intelligence and anthropomorphism?
- How does authorship attribution (AI alone vs. AI-human co-authorship) affect evaluations of both the content and the respective author?
- Does co-authorship mitigate potential negative effects of disclaimers emphasizing AI's limitations on credibility perceptions?

To answer these research questions, we designed a series of studies investigating the influence of different disclaimer types on credibility perceptions. In addition, we address two further aspects that can influence how an AI-generated text's credibility and the AI itself are assessed, that is information presentation and co-authorship.

1.2. The current research

First, we examined how disclaimers that highlight either the strengths or the limitations of the technology at work influenced content evaluations (Studies 1–3). Since there is little knowledge among users about the capabilities of AI systems (Lermann Henestrosa & Kimmerle, 2024), new information could influence current assessments of the tool. Moreover, investigating these spontaneous influences is crucial given the absence of legal regulations and the lack of control over how companies provide and emphasize information about their products.

Second, we considered the capabilities of LLMs to communicate information in various styles and manipulate information presentation by presenting information neutrally vs. evaluatively (Study 1). Previous studies have shown that an evaluative communication style can lower credibility ratings (Lermann Henestrosa et al., 2023; Tandoc et al., 2020). To investigate how the communication style interacts with the disclaimer type, we examined whether and how different information about an AI changes the subsequent evaluation of the AI and its output and whether this is influenced by presentation style.

Third, we examined whether co-authorship could balance information provided about the weaknesses of an AI and its potential (Studies 2 & 3). While pure AI authorship is relatively clear when it comes to generating output on demand, AI-generated texts are increasingly being copied and distributed elsewhere. Purely AI-authored production of an article for science journalism without at least human review and sign-off should be a rarity. Research on algorithm aversion has shown that a human-algorithm hybrid is evaluated more favorably than solely an algorithm (Palmeira & Spassova, 2015). Moreover, according to Jussupow et al. (2020), human involvement can increase the perceived capabilities of the algorithms and decrease their agency, which both can result in a more favorable perception. Therefore, this research aims to provide first insights into this interplay of disclaimers and different authorship conditions. Especially for areas where information must be checked to the best of one's knowledge and belief, and a human in the loop is virtually indispensable, it is highly relevant to investigate whether the human is perceived as a control authority compensating for the AI's potential flaws.

In all the following studies, science-related topics served as the context for AI communication. Using AI to retrieve and provide scientific information, is a particular case that offers the opportunity to look at the relationship between AI and science communication from different perspectives. While LLMs are designed to disclose when they lack data in certain areas, their basic principle—calculating the next most likely word—limits their reliability in science communication. Nevertheless, it is to be considered that AI will function as a novel source and mediator for communication, even in areas where its data is limited. Therefore, the focus should be on empowering people to verify this information instead of discouraging them from seeking it.

2. Study 1

The following experiment took up various approaches from previous studies by introducing a disclaimer that had an explanatory function and thus informed users about this new technology. The particular capabilities and weaknesses of LLMs were addressed to find out whether information about these has an impact on the perception of the output. In doing so, we also approached the question of whether disclaimers that inform about GenAI's boundaries or strengths are effective in terms of reducing or increasing credibility. Such limitations include the accuracy of the models when it comes to querying facts (OpenAI), but also systematic biases in the training data that arise, for example, from the underrepresentation of marginalized groups. Since users may not be aware of those issues, there is an increasing demand for the promotion of AI literacy. Previous research has revealed people's tendency to imagine computer processing as similar to human thinking (Sulmont et al., 2019, pp. 948–954) and that lower AI literacy is associated with greater receptivity to AI and perceiving it as "magical" (Tully et al., 2025). Therefore, it is necessary to investigate the effect of information about AI systems presented to users.

At the same time, we considered GenAI's capabilities to present information in various styles. For example, the same information can be presented in a neutral style or in an evaluative tone for purposes of emphasizing or contextualizing information. Information presentation has been found to influence perceived credibility, with evaluative language reducing it (Lermann Henestrosa et al., 2023). Therefore, we added information presentation as a repeated measurement to give participants the opportunity for a direct comparison between presentation styles and to assess whether information presentation interacts with disclaimer type.

Furthermore, we aimed to assess participants' general perceptions of AI vs. human authorship using the *machine heuristic* (Sundar & Kim, 2019, pp. 1–9). The machine heuristic comprises certain characteristics people can have in mind when being confronted with a machine, or here, with AI. According to the MAIN-model (Sundar & Kim, 2019, pp. 1–9), technological affordances, such as interface cues, can trigger this heuristic, which, in turn, can shape the perception of and interaction with technology. Indeed, previous research found the machine heuristic being activated when confronted with machines taking over human tasks (Molina & Sundar, 2024; Wang, 2021), and the predicted features can be specifically relevant in the evaluation of texts, like being accurate. Therefore, we asked participants to make the comparison between a human and an AI regarding those adjectives to exploratorily contrast their attributions toward both potential authorships. As we also wanted to capture a more general indication of whether participants intended to consume AI-written content on scientific information, we added two items based on the technology acceptance model (Venkatesh et al., 2003). We state the following hypotheses [1].

[1] The hypotheses were preregistered at https://aspredicted.org/jp93-dsy8.pdf; Note that we have changed the original terms used in the preregistration for the purpose of consistency in this paper as follows: We have renamed the factor *AI priming* to *disclaimer type*, and the levels *control* to *basic*, and *weaknesses* to *limitations*.

H.1.1a. We expected a main effect of disclaimer type on message credibility with higher ratings in the strengths condition compared to the limitations condition.

H.1.1b. We expected a main effect of information presentation on message credibility with higher ratings when the information is presented neutrally than when it is presented evaluatively.

H.1.1c. We expected an interaction effect of disclaimer type and information presentation on perceived message credibility with highest ratings in the strengths condition and neutrally presented information.

H.1.2a. We expected a main effect of disclaimer type on source credibility with higher ratings in the strengths condition compared to the limitations condition.

H.1.2b. We expected a main effect of information presentation on source credibility with higher ratings when the information is presented neutrally than when it is presented evaluatively.

H.1.2c. We expected an interaction effect of disclaimer type and information presentation on perceived source credibility with highest ratings in the strengths condition and neutrally presented information.

H.1.3. We expected a main effect of information presentation on anthropomorphism with higher ratings when the information is presented evaluatively than when it is presented neutrally.

2.1. Method

The study was approved by the ethical committee of the Leibniz-Institut fuer Wissensmedien in Tuebingen (LEK 2021/150).

2.1.1. Design and participants

In a 3 × 2 mixed-design, we manipulated the type of disclaimer (basic vs. strengths vs. limitations) as a between-group factor and information presentation (neutral vs. evaluative) as a within-group factor. An a priori power analysis determined a sample size of 342 participants to achieve 80 % power if alpha = 0.05 and $f = 0.10$ was assumed to be the minimum effect size of interest. A total of 657 participants were collected due to high exclusion rates: 32 participants were excluded because they withdrew their data at the end of the study and due to failed attention checks regarding the content of the articles or the authorship. Moreover, 284 participants who incorrectly answered the manipulation check questions regarding the disclaimer type were excluded. Thus, the total sample size resulted in $N = 341$. Table 1 shows the gender, age, and educational level distributions[2] for all three studies presented here. Participants were recruited via Prolific in November 2021. Participation took $M = 12.06$ min and was compensated with 2 Pounds Sterling. The study was conducted as an online survey designed on the Qualtrics platform and was only made accessible to German-speaking persons over 18 years old.

2.1.2. Material and procedure

Participants were randomly assigned to one of three introductory texts about AI that informed them about AI and automated text generation (ATG) in general or had additional information about either the strengths or the weaknesses of ATG. See Appendix A for the manipulation of the disclaimer type. Afterward, respondents were asked to read two science journalistic text excerpts that were presented in random order and allegedly written by AI. The articles both described the human impact on climate change and the possible consequences of global warming (see Appendix B). Information presentation was varied by presenting one article written in a neutral, objective way and the other article written evaluatively. To differentiate the evaluative article and emphasize the perceived subjectivity, emotional words like *catastrophic*, *terrifying*, *shocking*, *to be clearly criticized*, or *negative* were chosen. Both articles were similar in length (264 words for the neutral article; 270 words for the evaluative article) and contained the same facts.

2.1.3. Measures

Prior to the disclaimer type manipulation and the presentation of the articles, participants were asked to indicate if they were skeptical toward AI, if they were familiar with ATG, and if they had some knowledge about AI on three single items. The items were measured on a 5-point Likert scale from 1 = absolutely disagree to 5 = absolutely agree. Moreover, a 10-point bipolar item adapted from (Lombardi et al., 2013; Lombardi & Sinatra, 2012) was presented, asking how plausible participants perceived climate change to be human-made from 1 = extremely improbable or impossible to 10 = highly plausible.

As a manipulation check concerning the disclaimer type, we used a multiple-choice item asking participants to recall the representation of AI right after its display. The options provided were "the weaknesses were emphasized", "neutral", "the strengths were emphasized", or "I don't know". Moreover, participants were asked to recall the author of the texts, which was announced before text presentation as well as in a byline. The options provided for the authorship check were "a journalist", "an AI", "a robot", "a climate activist", or "I don't know".

After each article, participants were presented with the following set of measures: perceived neutrality of the text (one item on a 5-point scale from 1 = absolutely neutral to 5 = absolutely evaluative), perceived message credibility ($\alpha_{neu} = 0.81$, $\alpha_{eva} = 0.85$) and source credibility ($\alpha_{neu} = 0.86$, $\alpha_{eva} = 0.86$) from Flanagin and Metzger (2000; each on five items on a 7-point bipolar scale, e.g., "incredible/credible"), perceived anthropomorphism ($\alpha_{neu} = 0.86$, $\alpha_{eva} = 0.80$) and intelligence ($\alpha_{neu} = 0.81$, $\alpha_{eva} = 0.89$) of the AI, adopted from Bartneck et al. (2009; each on five items on a 5-point bipolar scale, e.g., "fake/natural" and "incompetent/competent"), belief in the machine heuristic regarding the AI ($\alpha_{neu} = 0.76$, $\alpha_{eva} = 0.77$) and regarding a human author ($\alpha_{neu} = 0.71$, $\alpha_{eva} = 0.71$) from Sundar and Kim (2019, pp. 1–9; each on four items measured on a 5-point Likert scale from 1 = absolutely disagree to 5 = absolutely agree, e.g., "If an AI/a human writes a text it/he writes [objective]"), and lastly behavioral intention ($\alpha_{neu} = 0.81$, $\alpha_{eva} = 0.89$; three items measured on a 5-point Likert scale from 1 = absolutely disagree to 5 = absolutely agree, e.g., "I would like to read a text like this again"). The individual items for each scale can be found in Appendix C.

2.2. Results

First, to check whether the manipulation of the information presentation was successful, we performed a *t*-test on the item regarding the perceived evaluation of the article. Indeed, the evaluative article was rated more evaluative ($M = 3.78$, $SD = 0.98$) than the neutral article ($M = 2.37$, $SD = 1.03$), $t(678.08) = 18.33$, $p < .001$ across all disclaimer type conditions. See Table 2 for the means and standard deviations of all dependent variables by condition.

Although we preregistered to calculate ANOVAs for Studies 1 and 2, we adopted regression models consistently across both studies, after carefully considering the data structure and hypotheses (Fox, 2016). This approach aligns with Study 3's preregistration and allows us to test our specific interaction hypotheses directly within a model instead of conducting multiple post-hoc tests as required by an ANOVA. It also improves statistical power and enhances interpretability of the effects regarding whether they support the hypotheses.

Therefore, a series of linear mixed-effects models with random intercepts for participants were conducted to examine the effects of disclaimer type and information presentation on the dependent variables in Study 1. Separate regression models were calculated for

[2] Educational levels were grouped into *low* (no degree, some high school or less, prefer not to say), *middle* (secondary school, vocational school, apprenticeship), and *high* (advanced technical college certificate, high school diploma, bachelor's degree, graduate or professional degree).

Table 1
Sample size, gender, age, and education levels for Studies 1-3.

		Gender			Age		Education		
	N	male	female	non-binary	*M* (*SD*)	range	low	middle	high
Study 1	341	129	202	10	28.07 (9.69)	18–67	1	47	293
Study 2	409	231	169	9	43.34 (14.53)	18–81	6	165	238
Study 3	512	234	277	1	46.45 (14.57)	18–69	129	158	225

Note: The option "diverse" in Study 2 includes "non-binary/third gender" and "prefer not to say".

Table 2
Means and standard deviations for all variables by information presentation and disclaimer type in Study 1.

	Neutral			Evaluative		
	strengths	limitations	basic	strengths	limitations	basic
Message credibility	5.90 (0.82)	5.65 (0.94)	5.79 (0.86)	5.01 (1.10)	4.85 (1.06)	4.64 (1.16)
Source credibility	5.85 (0.93)	5.59 (1.03)	5.77 (0.93)	4.92 (1.17)	4.90 (1.04)	4.58 (1.26)
Anthropomorphism	3.42 (0.90)	3.19 (0.88)	3.43 (0.86)	3.82 (0.64)	3.56 (0.78)	3.68 (0.78)
Intelligence	4.28 (0.57)	4.05 (0.63)	4.21 (0.55)	4.07 (0.65)	3.95 (0.78)	3.86 (0.80)
Behavioral intention	3.55 (0.90)	3.05 (0.88)	3.42 (0.90)	3.30 (0.87)	3.06 (0.93)	3.07 (0.85)
Item evaluation	2.30 (0.97)	2.37 (1.05)	2.49 (1.13)	3.66 (1.00)	3.79 (1.02)	4.00 (0.89)
Machine heuristic						
AI	3.75 (0.69)	3.47 (0.75)	3.52 (0.78)	3.63 (0.74)	3.43 (0.78)	3.4 (0.74)
Human journalist	2.96 (0.55)	2.88 (0.55)	2.90 (0.69)	3.01 (0.57)	2.90 (0.63)	2.82 (0.71)
Familiarity with ATG	2.74 (1.06)					
Self-assessed knowledge about AI	3.11 (0.92)					
Human made climate change plausibility	9.30 (1.24)					
Skepticism toward AI	2.73 (0.92)					

message credibility, source credibility, and anthropomorphism, with disclaimer type and information presentation as fixed effects and participant ID as a random effect.[3] Custom contrasts were set to explore the specific effects of the disclaimer types: (1, −1, 0) for comparing the strengths vs. the limitations disclaimer (hypotheses H.1.1a & H.1.2a), and (1, −0.5, −0.5) for assessing if the strengths disclaimer induces the highest ratings in the neutral condition (hypothesis H.1.1c & H.1.2c). As a first step, we compared for each dependent variable a model including only the main effects with a model incorporating the interaction of disclaimer type and information presentation. Only if the interaction contributed significantly to the model fit, as indicated by a likelihood ratio test ($p < .05$), the model with the interaction term was interpreted and reported in the following. The levels "strengths" and "evaluative" served as reference categories.

Contrary to H.1.1a, perceived message credibility was not higher after being presented with the strengths disclaimer than when being presented with the limitations disclaimer, $\widehat{\beta} = -0.04$, 95 % CI [−0.29, 0.22], $t(338) = -0.27$, $p = .784$. However, in line with H.1.1b, a significant effect of information presentation occurred, $\widehat{\beta} = 0.95$, 95 % CI [0.83, 1.06], $t(340) = 15.60$, $p < .001$, indicating higher message credibility ratings when information was presented neutrally than when presented evaluatively. The model with interaction term did not contribute significantly, thus not supporting H.1.1c. Fig. 1 depicts the means and distributions for message credibility by disclaimer type and information presentation.

Regarding H1.2a, the model revealed only a marginally significant difference between the strengths and limitations conditions, $\widehat{\beta} = -0.33$, 95 % CI [−0.66, 0.00], $t(600.63) = -1.95$, $p = .051$, but a significant effect of the strengths disclaimer compared to both other conditions, $\widehat{\beta} = 0.46$, 95 % CI [0.11, 0.81], $t(600.63) = 2.59$, $p = .010$, suggesting the highest ratings for source credibility when participants were presented with the strengths disclaimer. In line with hypothesis H.1.2b, a significant effect of information presentation was observed, with higher source credibility ratings when the information was presented neutrally, $\widehat{\beta} = 0.94$, 95 % CI [0.80, 1.07], $t(338) = 13.18$, $p < .001$. Moreover, both interaction terms became significant, suggesting a difference in the effect of the strengths compared to the limitations disclaimer depending on information presentation, $\widehat{\beta} = 0.53$, 95 % CI [0.15, 0.91], $t(338) = 2.74$, $p = .006$, as well as a difference in the effect of the strengths compared to both other disclaimer conditions depending on information presentation, $\widehat{\beta} = -0.54$, 95 % CI [−0.93, −0.14], $t(338) = -2.66$, $p = .008$. The pairwise comparisons revealed no significant differences in strengths compared to limitations ($p > .452$) as well as strengths compared to basic ($p > .088$) at both information presentation levels (see Fig. 2).

In line with H.1.3, the effect of information presentation on perceived anthropomorphism was significant, $\widehat{\beta} = -0.35$, 95 % CI [−0.45, −0.25], $t(340) = -6.91$, $p < .001$, revealing that the AI was perceived as significantly less anthropomorphic when the article was written neutrally. Furthermore, the effect of the strengths compared to the limitation disclaimer condition was marginally significant, $\widehat{\beta} = 0.18$, 95 % CI [−0.02, 0.38], $t(338) = 1.74$, $p = .083$. The anthropomorphism distributions across disclaimer type and information presentation are depicted in Fig. 3.

2.2.1. Further analyses

As we also measured the perceived intelligence of the AI author, behavioral intentions, and the belief in the machine heuristic for both an AI and a human journalist when writing a text, we exploratively examined if disclaimer type and information presentation affected these variables. The same contrasts as for the hypotheses were applied to the factor disclaimer type.

Regarding intelligence, the model revealed a significant effect of information presentation, with higher ascribed intelligence toward the AI when the information was presented neutrally, $\widehat{\beta} = 0.23$, 95 % CI [0.15, 0.30], $t(340) = 6.10$, $p < .001$. Moreover, participants indicated significantly higher behavioral intentions regarding the neutrally written article than for the evaluatively article, $\widehat{\beta} = 0.23$, 95 % CI [0.12, 0.34], $t(340) = 3.99$, $p < .001$.

For the belief in the machine heuristic, the model revealed a signif-

[3] The software used for the statistical analyses of all three studies was RStudio (version 4.4.2).

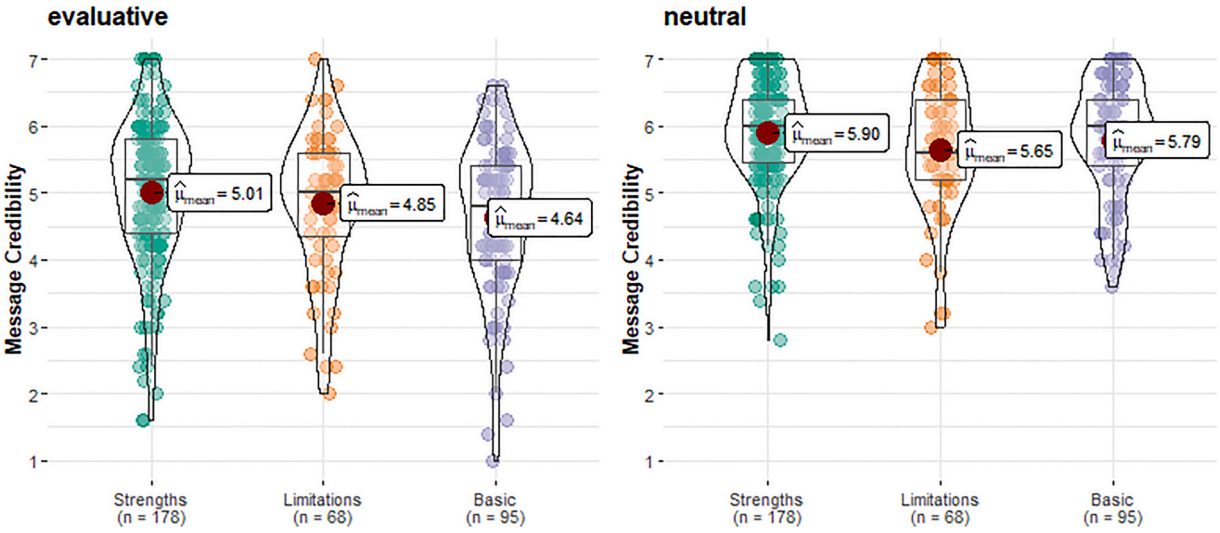


Fig. 1. Differences in perceived message credibility by disclaimer type and information presentation in Study 1.

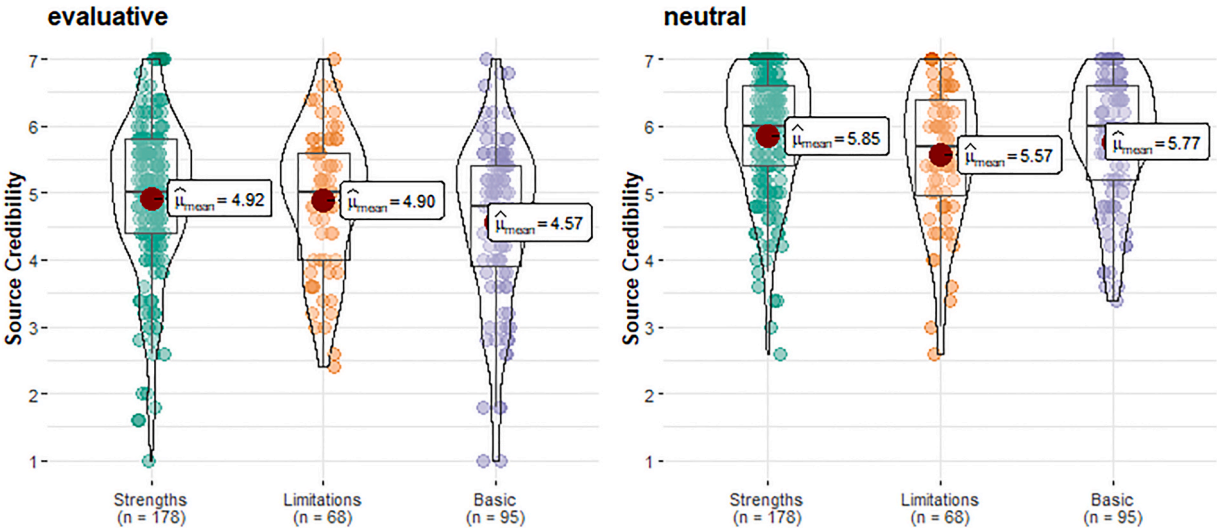


Fig. 2. Differences in perceived source credibility by disclaimer type and information presentation in Study 1.

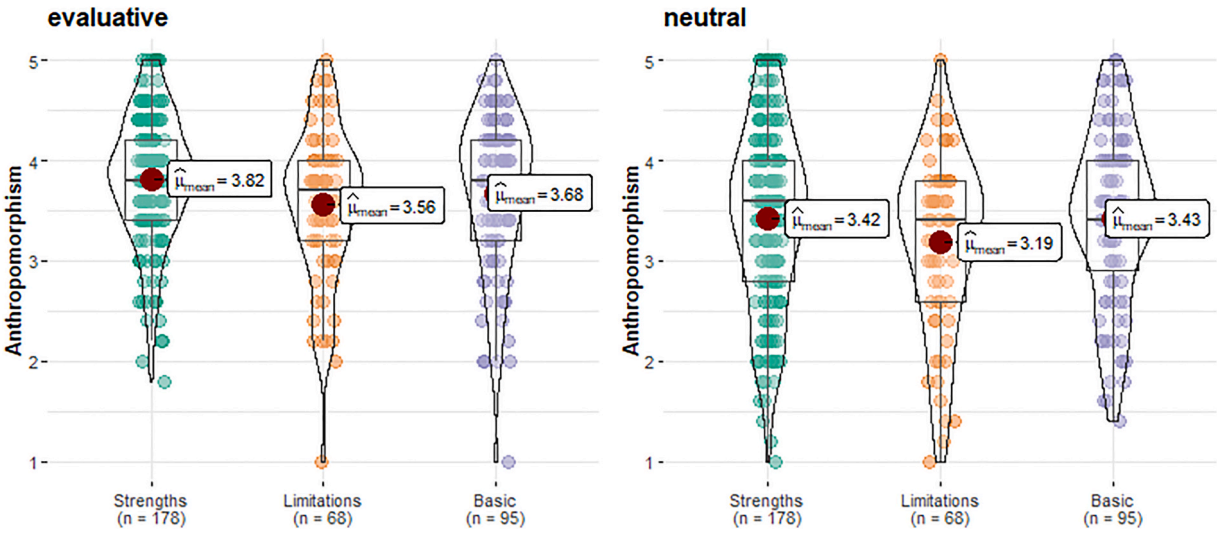


Fig. 3. Differences in perceived anthropomorphism by disclaimer type and information presentation in Study 1.

icant effect of authorship, $\widehat{\beta} = -0.65$, 95 % CI [−0.70, −0.60], $t(1021) = -24.85$, $p < .001$. Thus, in the direct comparison, participants rated a human journalist lower on the adjectives neutral, objective, accurate, and unbiased than an AI when writing a text. The effect of information presentation became marginally significant, $\widehat{\beta} = 0.05$, 95 % CI [−0.01, 0.10], $t(1021) = 1.77$, $p = .078$.

2.3. Discussion

This study replicated findings from Lermann Henestrosa et al. (2023), showing that an evaluative tone in an AI-written article increased perceived anthropomorphism while decreasing perceived credibility and intelligence. However, our hypotheses regarding the main effect of disclaimer type on credibility ratings, with the highest values when the strengths are presented, were only partly supported. While descriptively, the ratings were highest for the strengths condition regarding both source and message credibility, the strengths and limitations conditions did not differ significantly. However, the results indicate that disclaimer type could have effects, particularly regarding the credibility of the source, but also depending on how the information is presented. More research is needed to examine the interplay of those factors, for instance, under which circumstances the disclaimer information might operate better or worse.

However, the absence of major differences combined with the overall high credibility ratings, especially in the neutral condition, could be interpreted as such that prior information, even if it highlights the pitfalls of technology, barely influences the evaluation of a subsequent output. To this effect, participants' tendency to ascribe the features of the machine heuristic more to an AI than to a human as well as the relatively high ascribed intelligence ratings, are evidence that there is already a basic confidence in the technologies' abilities. This raises the question of how a certain type of authorship in connection with the respective disclaimer information affects the recipients' perception, specifically if the presence of a human in the authorship attribution interacts with the information presented in the disclaimers. Since the manipulation check regarding disclaimer type also revealed that participants perceived already the basic disclaimer to be an emphasis of the strengths of the AI, Study 2 addressed this issue by emphasizing the limitations and biases associated with LLMs more strongly.

3. Study 2

Study 2 explored the interaction between emphasizing an AI's limitations and the presence of combined authorship by a human and AI (co-authorship). Disclaimer types outlining system capabilities serve as a direct means of informing users about issues of technology. However, the results of the machine heuristic in Study 1 suggested that people may also hold beliefs regarding the potential disadvantages of human authorship. Although professional research practices and references to various sources are integral to journalistic standards, the opinion expressed in an article can sometimes reflect the author's subjective viewpoint. Conversely, the limitations of AI could be mitigated by including a human co-author if the reader considers them as a verifier. Therefore, this study addressed the prevalent practice of human-AI co-authorship and examined the effects of highlighting only the AI's limitations (AI limitations condition) or also addressing those of human authors (AI & human limitations condition). We state the following hypothesis[4].

[4] The hypotheses were preregistered at https://aspredicted.org/vvrp-md4p.pdf; Note that we have changed the original terms used in the preregistration for the purpose of consistency in this paper as follows: we have renamed the factor *bias information* to *disclaimer type* and renamed the levels *basic AI information* to *basic*, *AI bias information* to *AI limitations*, and *AI + human bias information* to *AI & human limitations*.

H.2.1a. We expected a main effect of disclaimer type on perceived message credibility with higher ratings in the basic disclaimer condition than in the AI limitations condition.

H.2.1b. We expected an interaction effect of disclaimer type and labeled authorship on perceived message credibility, with higher ratings in the AI limitations condition when the labeled authorship is human co-authorship compared to sole AI authorship.

H.2.2a. We expected a main effect of disclaimer type on perceived source credibility with higher ratings in the basic disclaimer condition than in the AI limitations condition.

H.2.2b. We expected an interaction effect of disclaimer type and labeled authorship on perceived source credibility with higher ratings in the AI limitations condition when the labeled authorship is human co-authorship compared to sole AI authorship.

H.2.3. We expected a main effect of labeled authorship on perceived intelligence with higher ratings for co-authorship than for AI authorship.

3.1. Method

Study 2 was approved by the ethical committee of the Leibniz-Institut fuer Wissensmedien in Tuebingen (LEK 2022/032).

3.1.1. Design and participants

The experimental design was a 3 × 2 between-groups design with *disclaimer type* (basic vs. AI limitations vs. AI & human limitations) and *labeled authorship* (AI vs. Co-authorship) as the two factors. According to an a priori power analysis, a total sample size of 432 participants was needed to detect a minimum effect size of $f = 0.15$ if alpha = 0.05 and a power of 80 %. To account for potential exclusions, 554 participants were recruited, of which six withdrew their data at the end of the experiment. Due to predefined exclusion criteria, that is, the manipulation and attention checks regarding the disclaimer type, the authorship, and the content of the article, another 145 participants had to be excluded from the analysis. Therefore, the total sample size resulted in $N = 409$. See Table 1 for demographic distribution.

A US sample over 18 years old was recruited via Prolific in July 2022. The online experiment had a mean duration of $M = 14.98$ min and was compensated with 2.28 Pound Sterling.

3.1.2. Material and procedure

Participants were randomly assigned to one of the six conditions resulting from the between-groups design. Hence, every participant was presented with one disclaimer type (see Appendix D) and afterward read an article about fat-shaming allegedly written by an AI author or co-authorship. Therefore, the article was identical for all respondents and differed only in the labeled authorship, introduced above the text and as a byline under the text as follows*: "This article was written by an automated text generator" for the AI author condition and "This article was written by an automated text generator, and checked and revised by a human" for the co-author condition.* The text was 537 words long and contained information about obesity and the negative consequences of fat-shaming and stereotypical thinking. The destructive effects of fat-shaming were communicated in an educational and evaluative style. All information included in the article was retrieved from reliable sources which were also listed after its final paragraph. See Appendix E for the full text.

3.1.3. Measures

Before being presented with the disclaimer type manipulation, respondents filled in the fat-phobia scale (Bacon et al., 2001, p. 14 bipolar items measured on a 5-point scale, e.g., "lazy – industrious"), which was again measured at the end of the experiment ($\alpha_{pre} = 0.93$, $\alpha_{post} = 94$). After the disclaimer, participants were asked whether it was true or false

that they had been informed about ATG (after the basic disclaimer), about ATG and the underlying risks (after the AI limitations disclaimer), or about ATG, its underlying risks, and biases within human articles (after the AI & human limitations disclaimer). Furthermore, a manipulation check regarding the authorship was presented after the article, including the options "automated text generator", "human journalist", "automated text generator and human", and "medical student". Afterward, the following measures followed: perceived neutrality of the text (one item on a 5-point scale from 1 = absolutely neutral to 5 = absolutely evaluative), perceived message credibility (α = .94; Appelman & Sundar, 2016; Sundar, 1999, p. 22 items measured on a 7-point Likert scale, e.g. "objective"), source credibility (α = .92; Flanagin & Metzger, 2000; five items on a 7-point bipolar scale), perceived anthropomorphism (α = .91) and intelligence (α = .93; Bartneck et al., 2009; each on five items on a 5-point bipolar scale), and behavioral intentions (α = .89; two items on a 5-point Likert scale, e.g., "I would read such an article again"). All scales used in Study 2 and their corresponding items can be found in Appendix F.

3.2. Results

Table 3 shows the means and standard deviation of all variables by condition.

Linear regressions with disclaimer type and labeled authorship as predictors were performed across all dependent variables. The AI author condition served as a reference category for the factor labeled authorship (dummy coded). Regarding the factor disclaimer type, custom contrasts were defined to explore the specific effects of the different disclaimers. (1, −1, 0) to compare the basic disclaimer vs. the AI limitations disclaimer (hypotheses H.2.1a and H2a) as well as the interaction of this comparison with labeled authorship (H.2.1b, H.2.2b), and (−0.5, −0.5, 1) to compare AI & human limitations vs. the other two disclaimer types (RQ2). Models including the interaction of disclaimer type and authorship will only be reported if the interaction term contributes significantly to the model fit. We did not define hypotheses regarding the main effect of authorship. As we state an open research question concerning such effect and it is part of the interaction hypotheses, however, we will include this factor as fixed effect in the regression models and exploratorily examine its effect.

Hypothesis H.2.1a predicted higher ratings for message credibility when participants were presented with the basic disclaimer compared to when presented with the AI limitations disclaimer. However, the model revealed only a marginal effect of this comparison, $\widehat{\beta} = 0.13$, 95 % CI [−0.01, 0.26], $t(403) = 1.79$, $p = .075$. Moreover, no significant difference between the AI & human limitations disclaimer compared to the other two disclaimers occurred, $\widehat{\beta} = -0.06$, 95 % CI [−0.17, 0.16], $t(403) = -0.07$, $p = .946$, which speaks against a main effect of this condition. The interaction effect between the comparison of the basic and the AI limitations disclaimer and labeled authorship was significant, $\widehat{\beta} = -0.28$, 95 % CI [−0.51, −0.05], $t(403) = -2.39$, $p = .017$. Fig. 4 suggests that this interaction might stem from higher message credibility in the basic condition than in the AI limitations condition when an AI is an alleged author, but a reverse pattern when authorship is attributed to a co-author. Although the interaction was significant, pairwise comparisons revealed no significant differences between the levels of the interaction ($p > .475$). No further effects occurred.

Regarding source credibility, the model revealed a significant effect of the basic disclaimer compared to the AI limitations disclaimer, $\widehat{\beta} = 0.41$, 95 % CI [0.14, 0.68], $t(403) = 2.96$, $p = .003$. As this effect is positive, it contradicts H.2.2a. However, regarding H.2.2b, the model also revealed an interaction effect between the comparison of the basic and the AI limitations condition and labeled authorship, $\widehat{\beta} = -0.71$, 95 % CI [−1.16, −0.26], $t(403) = 3.11$, $p = .002$, with lower source credibility in the AI limitations condition only when the article was AI-authored (see Fig. 5). Indeed, the pairwise comparison revealed a significant difference between the basic and the AI limitations disclaimer types in the AI-authorship condition ($p = .038$) but not in the co-authorship condition ($p = .562$). Furthermore, the levels of the AI limitations condition by labeled authorship did not differ significantly ($p = .484$), which contradicts the assumption of H.2.2b. The AI & human limitations disclaimer did not differ significantly from the other disclaimer types, $\widehat{\beta} = -0.03$, 95 % CI [−0.28, 0.35], $t(403) = 0.19$, $p = .851$, thus speaking against a main effect of this conditionNo further effects occurred.

H.2.3 predicted a main effect of labeled authorship on perceived intelligence with lower ratings for AI-authorship, which the regression model did not support, $\widehat{\beta} = -0.03$, 95 % CI[−0.18, 0.13], $t(403) = -0.36$, $p = .718$. An exploratory examination of the effect of disclaimer type in perceived intelligence revealed a marginally significant difference between the basic disclaimer compared to the AI limitations disclaimer, $\widehat{\beta} = 0.11$, 95 % CI [−0.01, 0.22], $t(403) = 1.82$, 43 $p = .070$, and no significant effect of the AI & human limitations compared to the other disclaimer conditions, $\widehat{\beta} = 0.03$, 95 % CI [−0.10, 0.17], $t(403) = 0.49$, $p = .628$. Again, the interaction effect of the comparison between the basic and the AI limitations condition by authorship was significant, $\widehat{\beta} = -0.24$, 95 % CI [−0.43, −0.05], $t(403) = -2.49$, $p = .013$, indicating higher ratings for the basic disclaimer than for the AI limitations disclaimer in the AI-author condition and the opposite pattern in the co-author condition. However, the pairwise comparisons revealed no significant differences between the levels of the interaction ($p > .439$).

An exploratory analysis of the influence of the two factors on perceived anthropomorphism and behavioral intentions revealed no significant effects.

3.2.1. Further analysis

A model predicting the fat phobia score (dummy coded for the factor disclaimer type with basic as a reference) revealed a significant effect of the repeated measure, $\widehat{\beta} = -0.19$, 95 % CI [−0.24, −0.15], $t(408) = -8.50$, $p < .001$, indicating lower ratings and thus less fat phobia after having read the article. Moreover, there was a significant effect of the AI & human limitations disclaimer, $\widehat{\beta} = 0.19$, 95 % CI [0.06, 0.32], $t(405) = 2.80$, $p = .005$, suggesting higher overall fat phobia scores in this condition.

Table 3
Means and standard deviations for all variables by disclaimer type and labeled authorship in Study 2.

	AI-authorship			Co-authorship		
	basic	AI limitations	AI & human limitations	basic	AI limitations	AI & human limitations
Message credibility	5.35 (0.97)	5.10 (1.04)	5.22 (0.84)	4.96 (0.91)	5.27 (0.88)	5.27 (0.81)
Source credibility	6.32 (1.89)	5.51 (1.75)	5.96 (1.75)	5.45 (1.61)	6.05 (1.88)	5.88 (1.75)
Anthropomorphism	3.80 (0.93)	3.62 (0.91)	3.72 (0.87)	3.60 (0.87)	3.79 (0.93)	3.83 (0.73)
Intelligence	4.27 (0.79)	4.06 (0.82)	4.21 (0.68)	4.03 (0.73)	4.30 (0.77)	4.13 (0.79)
Behavioral intentions	3.63 (1.18)	3.36 (1.23)	3.51 (0.97)	3.11 (1.12)	3.31 (1.12)	3.46 (1.01)
Fat-phobia pre	3.65 (0.58)	3.70 (0.60)	3.76 (0.64)	3.60 (0.61)	3.72 (0.54)	3.80 (0.49)
Fat-phobia post	3.40 (0.66)	3.51 (0.60)	3.63 (0.62)	3.43 (0.66)	3.48 (0.62)	3.66 (0.55)
Item evaluation	3.80 (0.90)	3.58 (0.93)	3.65 (0.87)	3.78 (0.73)	3.65 (1.04)	3.68 (0.87)

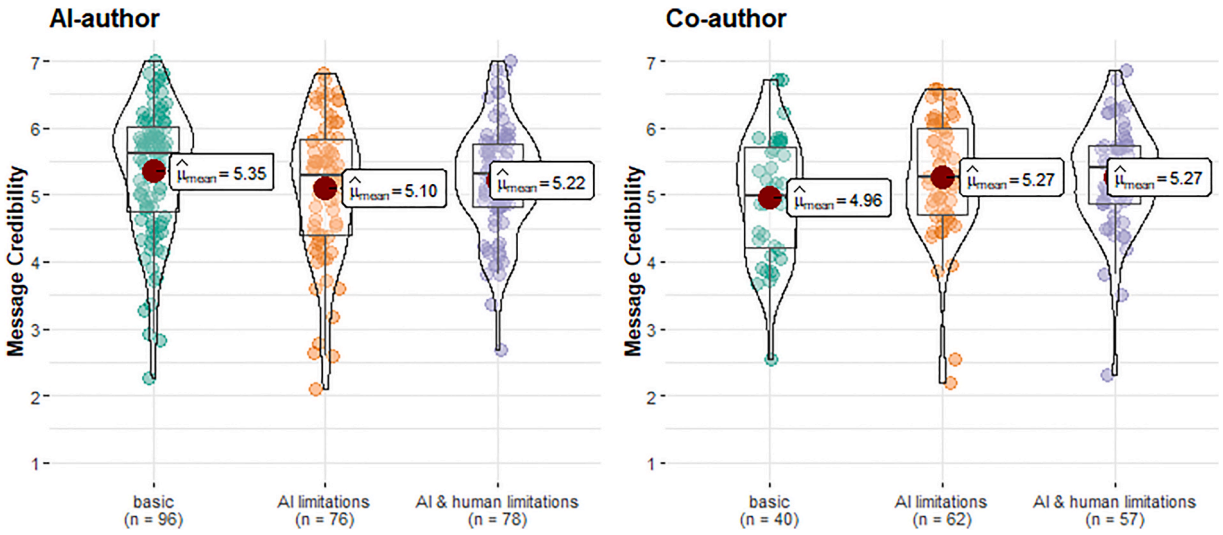


Fig. 4. Differences in perceived message credibility by disclaimer type and labeled authorship in Study 2.

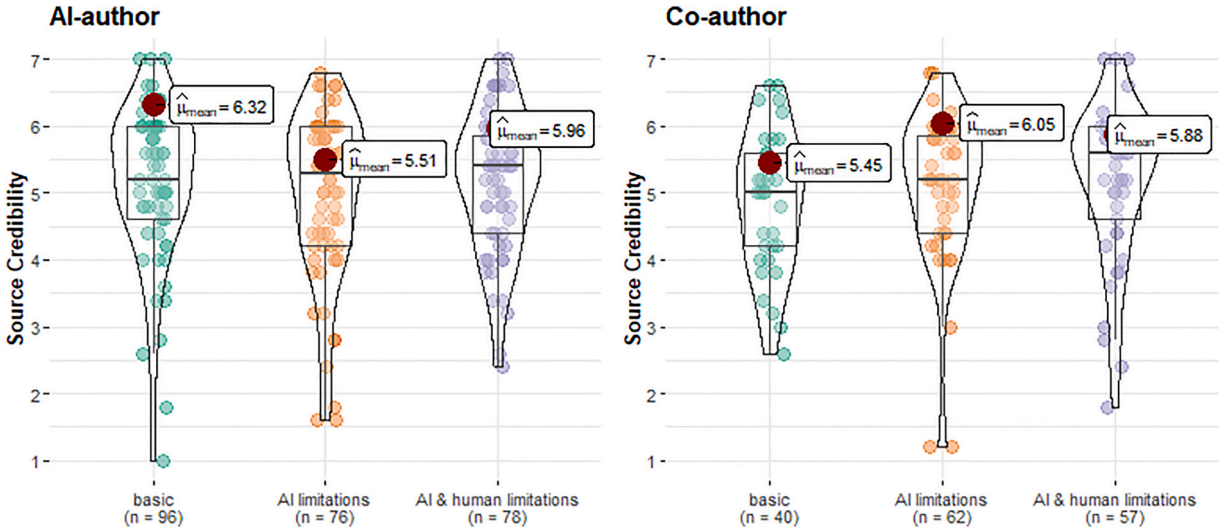


Fig. 5. Differences in perceived source credibility by disclaimer type and labeled authorship in Study 2.

3.3. Discussion

Presenting participants with information about the limitations of an AI or additionally human flaws compared to a basic disclaimer had no negative effect on message credibility ratings. The significant interaction suggests different patterns depending on authorship attribution, which require further investigation with balanced and sufficiently large sample sizes. Regarding source credibility, the results are rather ambiguous. The significant interactions suggest that AI authorship induced more positive credibility ratings than co-authorship, but only when a basic definition was presented beforehand. Though the pairwise comparisons mostly revealed no significant differences between the levels of the interactions, the result could be an indication of a balancing effect of a human involved when limitations are highlighted as well as a balancing effect of also emphasizing human flaws in the case of sole AI authorship. The decrease in the fat phobia score after reading an article about its destructive effects suggests a significant influence of the article's content on participants' attitudes. However, this decrease could also stem from a social desirability bias and would require a repeated measurement after some time to check whether it is lasting. Concerning the perceived intelligence and anthropomorphism of the authors, the basic assumption that the authors would differ here must be rejected.

Overall, the results must be interpreted cautiously due to the imbalanced group sizes and high dropouts in the co-author condition. This can be seen because of the manipulation check design, which led to high exclusion rates and unbalanced group sizes. Moreover, in this study, we added the references used at the bottom of the article, which might have increased credibility ratings overall and do not reflect common LLMs, which often fail to provide accurate references if they do so. Finally, presenting people with certain information about ATG without controlling for prior knowledge or attitudes toward this technology, interfering effects of these factors cannot be ruled out.

In sum, our results speak against major effects of disclaimer type, especially on text perception. Whether the presentation of any information in general already inspires confidence, however, can only be examined in comparison with a condition without prior information. Therefore, we addressed the relationship between disclaimer type and co-authorship in an adapted design in the third study to fully map the variations of disclaimer types and to disentangle possible balancing or canceling conditions.

4. Study 3

In their meta-analysis, Wang and Huang (2024) found that in the presence of system transparency, that is, when informing about the algorithm, machine-written content induced lower credibility ratings compared to human-written content. To address this lack of a condition without a disclaimer and, thus, no system transparency, we included such a condition in this experiment. Ideally, readers should be informed about both the limitations and capabilities of LLMs so they can use them effectively while remaining critically aware of the potential for misinformation. Therefore, a fifth condition was added to this experimental design, comprising the limitations as well as the strengths in the disclaimer and thus providing balanced information to readers. Labeling authorship was an additional factor to further investigate the relationship between disclaimer type and authorship. In Study 2, co-authorship was introduced as an automatically generated text that was checked and revised by a human. This time, we emphasized the collaboration of a human and an LLM with a focus on the human as the principal author by describing co-authorship as a human writer who was being supported by an AI system. We state the following explorative research question and hypotheses[5].

RQ1: Is there a difference between the no information framing condition and the basic information condition?

If these two conditions do not differ, we will merge them in the subsequent analysis.

H.3.1a. We expected a main effect of authorship on message credibility, with higher ratings for articles allegedly co-authored compared to those authored solely by AI.

H.3.1b. We expected a main effect of disclaimer type on message credibility, with significantly higher ratings for articles in the strengths condition compared to the articles in other disclaimer conditions.

H.3.1c. We expected an interaction effect of authorship and disclaimer type on message credibility. Specifically, ratings will be significantly lower for the AI author than for the co-author in the limitations condition, compared to the other disclaimer conditions.

H.3.2a. We expected a main effect of authorship on source credibility, with higher ratings for articles allegedly co-authored compared to those authored solely by AI.

H.3.2b. We expected a main effect of disclaimer type on source credibility, with significantly higher ratings for articles in the strengths condition compared to the articles in other disclaimer conditions.

H.3.2c. We expected an interaction effect of authorship and disclaimer type on source credibility. Specifically, ratings will be significantly lower for the AI author than for the Co-author in the limitations condition, compared to the other disclaimer conditions.

H.3.3a. We expected a main effect of authorship on perceived anthropomorphism, with significantly higher ratings for the Human-AI-Co-author compared to the AI-author.

H.3.3b. We expected an interaction effect of authorship and disclaimer type on perceived anthropomorphism. Specifically, ratings will be significantly lower for the AI author than for the Co-author in the limitations condition, compared to the other disclaimer conditions.

H.3.4a. We expected a main effect of authorship on perceived intelligence, with significantly higher ratings for the Human-AI-Co-author compared to the AI-author.

H.3.4b. We expected an interaction effect of authorship and disclaimer type on perceived intelligence. Specifically, ratings will be significantly lower for the AI author than for the Co-author in the limitations condition, compared to the other disclaimer conditions.

H.3.5. We expected a main effect of authorship on behavioral intentions, with significantly higher ratings in the Human-AI-Co-author conditions compared to the AI-author conditions.

We state no further hypotheses regarding behavioral intentions, intelligence, and anthropomorphism. As we always test the full model, these effects were exploratively tested.

4.1. Method

This study was approved by the ethical committee of the Leibniz-Institut fuer Wissensmedien in Tuebingen (LEK 2022/032).

4.1.1. Design and participants

In a 5 × 2 between-groups design, we again investigated the effects of disclaimer type (no information vs. basic vs. strengths vs. limitation vs. balanced) and labeled authorship (AI vs. co-authorship). We calculated a total sample size of 614 participants to be able to detect an effect of $f = 0.14$, with alpha = 0.05 and a power of 80 %. Of 1031 participants who completed the experiment, 71 withdrew their data, and another 295 were excluded because of failed attention checks regarding the content of the article or failed manipulation checks regarding the authorship. From the remaining 665 participants, we excluded 154 respondents who incorrectly answered the recall test concerning the disclaimer type, resulting in a total sample size of $N = 512$ participants. The gender, age, and educational distributions are depicted in Table 1.

Participants were recruited via the panel provider Bilendi & respondi in August 2024. The sample consisted of German-speaking adults over 18 years old. The experiment had a mean duration of $M = 9.53$ and was compensated via the panel provider.

4.1.1.1. Material and procedure. Participants were assigned to one of ten conditions, thus seeing one out of five disclaimer types (note that in the no information condition, no disclaimer was shown; see Appendix G for the disclaimer types) before reading the text either allegedly written by AI or by human-AI-co-authorship. The AI was introduced as follows: "In the following, you will read a text that was written by an artificial intelligence, more precisely: a Large Language Model (LLM)". The co-authorship was introduced as follows: "In the following, you will read a text that was written by a human with the help of artificial intelligence, more precisely: a Large Language Model (LLM)". Additionally, a byline above the article again stated the authorship. The article was actually written by GPT 4 using the prompt "Write a short science journalistic article (continuous text without subheadings or bullet points), as it could appear in a German newspaper, on the topic 'The influence of social media use on the mental health of adults'" (originally German). The resulting output was 261 words long and contained information about the psychological consequences of social media use (see Appendix H).

4.1.2. Measures

First, participants were asked about their experience with AI-generated texts (if they have ever heard about this technology or read an AI-generated text, and if they have ever used an LLM) as well as their prior attitudes toward LLMs by five items (α = .89) adapted from Venkatesh et al. (2003; e.g., "I think AI-generated texts are a useful innovation"). Afterward, we measured self-assessed knowledge about AI on a 5-point scale from "no knowledge about AI" to "extensive knowledge about AI". After the disclaimer type was presented, a recall test tailored to each manipulation followed, asking participants to identify two out of four statements which were mentioned in the disclaimer. Thus, for the limitations condition, a correct statement was, for instance: "It was

[5] The hypotheses were preregistered at https://aspredicted.org/zyfb-mv8t.pdf; Note that we have renamed the factor *AI framing* to *disclaimer type*.

mentioned that LLMs have no real understanding". The two distractors were the same for all recall checks. Subsequently, the article appeared, followed by a set of measures: perceived neutrality of the text (one item on a 5-point scale from 1 = absolutely neutral to 5 = absolutely evaluative), perceived message credibility (α = .93; Sundar, 1999, p. 19 items measured on a 7-point Liker scale), source credibility (α = .91; Flanagin & Metzger, 2000; five items on a 7-point bipolar scale) perceived anthropomorphism (α = .89) and intelligence (α = .91; Bartneck et al., 2009; each on five items on a 5-point bipolar scale), and behavioral intentions (four items on a 5-point Likert scale). Please see Appendix I for the individual items per scale.

4.2. Results

Table 4 displays the means and standard deviations for all variables by condition. Since experience with LLMs and self-assessed knowledge about AI were queried on ordinal scales, the relative frequencies of each level are presented in Table 5.

We preregistered that we would first check if the disclaimer levels no information and basic information differed statistically. In fact, across all dependent variables, there were no significant differences between these two disclaimer types, so those groups were merged in the first step. Linear regression models with the remaining four disclaimer types and labeled authorship as fixed effects were calculated for each dependent variable. While the AI served as a reference category for the factor authorship, custom contrasts were defined for the factor disclaimer type to explore the specific effects of the different disclaimers. Thus, the contrasts were coded as follows: (−1/3, 1, −1/3, −1/3) for comparing the strengths disclaimer against the other disclaimers (hypotheses 3.1b and 3.2b), (−1/3, −1/3, 1, −1/3) for comparing the limitations against the other disclaimers in the interaction with authorship (H.3.1c and H.3.2c), and (0, 1, 0, −1) to examine exploratorily if the balanced disclaimer differed from the strengths. Again, the model including the interaction term will only be presented if it contributed significantly to the model fit.

Regarding message credibility, the interaction term did not contribute significantly to the model fit. H.3.1a predicted higher ratings of perceived message credibility for texts co-authored compared to sole AI-authored. However, no effect of authorship occurred, $\widehat{\beta} = 0.02$, 95 % CI [−0.14, 0.18], $t(507) = 0.30$, $p = .766$. Moreover, neither was the effect of strengths predicted in H.3.1b significant, $\widehat{\beta} = 0.16$, 95 % CI [−0.11, 0.43], $t(507) = 1.15$, $p = .253$, nor was there a significant difference between the strengths and the balanced disclaimer, $\widehat{\beta} = -0.09$, 95 % CI [−0.32, 0.15], $t(507) = -0.73$, $p = .468$.

For source credibility, the model with interaction term did also not contribute significantly to the model. Again, contradicting H.3.2a, there was no effect of authorship, $\widehat{\beta} = 0.10$, 95 % CI [−0.12, 0.31], $t(507) = 0.87$, $p = .385$. However, the effect of strengths predicted in H.3.2b was significant, $\widehat{\beta} = 0.38$, 95 % CI [0.01, 0.75], $t(507) = 2.04$, $p = .042$, indicating higher source credibility ratings after participants were presented with the strengths disclaimer. Furthermore, the effect of the limitations condition compared to the other conditions became significant, $\widehat{\beta} = 0.28$, 95 % CI [0.07, 0.50], $t(507) = 2.59$, $p = .010$, suggesting also higher perceived source credibility after being presented with the limitations disclaimer. The exploratorily examined difference of strengths versus balanced revealed no effect, $\widehat{\beta} = -0.12$, 95 % CI [−0.43, 0.20], $t(507) = -0.73$, $p = .465$. Fig. 6 depicts the distributions of source credibility by disclaimer type and labeled authorship.

Regarding anthropomorphism, the interaction did not contribute significantly, contradicting H.3.3b. However, the effect of labeled authorship became significant, $\widehat{\beta} = 0.16$, 95 % CI[0.03, 0.30], $t(507) = 2.34$, $p = .020$, supporting H.3.3a that the co-authorship was perceived to be more anthropomorphic.

The interaction term did not significantly contribute to the model predicting perceived intelligence, thus not supporting H.3.4b. H.3.4a

Table 4
Means and standard deviations of the variables by disclaimer type and labeled authorship in Study 3.

	AI-author					Co-author				
	no information	basic	strengths	limitations	balanced	no information	basic	strengths	limitations	balanced
Message credibility	5.01 (0.89)	5.01 (0.87)	5.14 (0.91)	5.35 (0.72)	5.15 (1.04)	4.94 (0.86)	5.44 (1.09)	5.25 (0.75)	5.12 (0.95)	5.16 (1.05)
Source credibility	5.25 (1.23)	5.15 (1.33)	5.54 (1.19)	5.66 (0.95)	5.24 (1.60)	5.06 (1.28)	5.67 (1.31)	5.73 (0.99)	5.59 (1.22)	5.47 (1.28)
Anthropomorphism	3.73 (0.83)	3.6 (0.83)	3.83 (0.71)	3.92 (0.67)	3.66 (0.99)	3.76 (0.80)	4.09 (0.68)	4.05 (0.64)	3.83 (0.78)	3.93 (0.79)
Intelligence	4.10 (0.70)	4.02 (0.74)	4.18 (0.68)	4.19 (0.72)	4.05 (0.95)	4.01 (0.65)	4.21 (0.80)	4.34 (0.63)	4.13 (0.73)	4.24 (0.64)
Item evaluation	2.53 (1.03)	2.5 (0.86)	2.35 (1.05)	2.41 (1.06)	2.85 (1.23)	2.63 (1.08)	2.76 (1.10)	2.62 (1.19)	2.39 (1.02)	2.25 (0.84)
Behavioral intention	3.00 (0.87)	2.96 (1.01)	3.32 (0.92)	3.29 (0.88)	3.21 (1.05)	3.14 (0.84)	3.32 (0.94)	3.21 (1.07)	3.05 (0.93)	3.23 (0.91)
Attitude toward LLMs	3.03 (0.38)									

Table 5
Absolute and relative frequencies (in relation to the total sample) of the ordinal scaled measures experience with LLMs and self-assessed knowledge about AI.

	Yes	No	Not sure
Heard that AI can write texts	483 94.34 %	29 5.66 %	–
Read an AI-written text	230 44.92 %	71 13.87 %	182 35.55 %
Ever used LLM(s)	158 30.86 %	72 69.14 %	–

	rarely	several times per month	once a week	several times per week	daily
LLM usage frequency	68 13.28 %	44 8.59 %	15 2.93 %	25 4.88 %	6 1.17 %

	no	little	some	considerable	extensive
Self-assessed knowledge	41 8.01 %	202 39.45 %	194 37.89 %	71 13.87 %	4 0.78 %

Note: Regarding the experience measures, the participants were only forwarded if they answered yes to the respective question, resulting in smaller subsamples for "read an AI-written text" and "ever used LLM(s)".

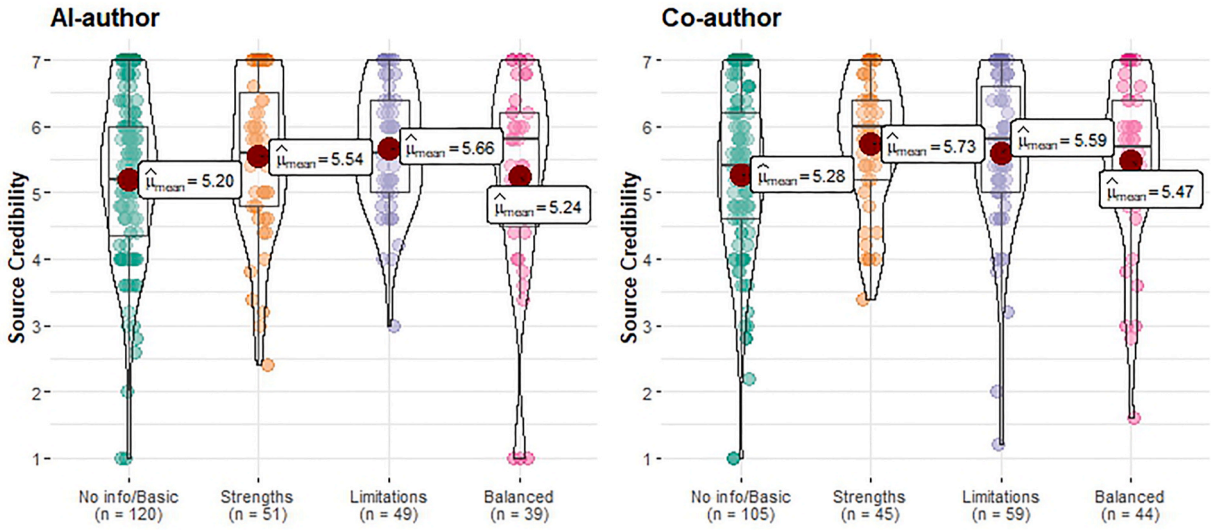


Fig. 6. Differences in perceived source credibility by disclaimer type and labeled authorship in Study 3.

predicted a main effect of authorship on the perceived intelligence of the author with higher ratings for Human-AI-co-authorship, which was not supported by the model, $\hat{\beta} = 0.06$, 95 % CI [−0.07, 0.18], $t(507) = 0.88$, $p = .382$. The effect of the strengths disclaimer compared to the other disclaimer conditions became only marginally significant, $\hat{\beta} = 0.20$, 95 % CI [−0.02, 0.41], $t(507) = 1.82$, $p = .069$. No significant effects occurred in the exploratory analysis of behavioral intentions.

4.3. Discussion

The results of Study 3 speak for hardly any effects of different disclaimer types on message credibility and for only small effects on source credibility perceptions. Therefore, the results are in line with Study 2, where neither a main effect of authorship nor negative effects of highlighting the pitfalls of AI occurred. Interestingly, both the strengths and the limitations conditions induced the highest source credibility ratings overall. In this regard, Dzindolet et al. (2003) showed that participants who were provided with an explanation of why a system might err showed higher trust and reliance in the decision aid, even when it was unwarranted. One assumption is, therefore, that additional information about a technology, even if it reveals weaknesses, already acts as a cue for credibility and trustworthiness, especially when people still know very little about the detailed functionality of such a system. However, it must be clearly questioned how much attention is paid to such information at all or if the ratings of the texts and the author rather depend on more persistent factors such as attitude toward AI and LLMs and the output's content. Moreover, contradictory to Wang and Huang (2024), who found lower credibility perceptions in the presence of explanatory information, our results speak against a difference between the no information condition and the basic information condition.

Furthermore, we found only an authorship effect regarding perceived anthropomorphism, which, on the one hand, suggests that slight differences between the authors were perceived, but these did not affect the evaluation of the text. Overall, perceived credibility and intelligence were rather high, again suggesting fundamental confidence in LLMs' capabilities. Nonetheless, it must be noted that an authorship label does not appear to be the driving influencing factor for the evaluation of texts. Above all, text characteristics, preconceptions about the topic, and about ATG in general should be examined as predictors. In this regard, the behavioral intentions to consume such a text again speak against a major interest in the topic but also in the use of LLMs in this sample. If cognitive motivation or interest in the topic moderate potential effects, is still to be investigated, especially as we did not assess prior attitudes in this study.

Again, as in Studies 1 and 2, the results must be interpreted considering the high dropout rates. The final sample size again fell behind the initially targeted one. To assess whether this impacts the interpretability of our findings, we conducted a sensitivity analysis, which indicated that our obtained sample size had sufficient power to detect effects of $f = 0.15$. Considering the overall small effects, the studies speak for a small influence of disclaimer type on text and author perceptions, if any. Nevertheless, we want to emphasize that the recall test used in this study shows that little attention is paid to disclaimers even in a controlled setting and that specific information is not retained for long. How more attention can be drawn to information provided by disclaimers and thus possibly generate stronger effects is a matter for future studies.

5. General discussion

GenAI, particularly LLMs, has the potential to revolutionize tasks involving natural language. Thus, agents and tools powered by these models, designed to respond to nearly any question, are being used by people for exactly this purpose—regardless of whether this is the most sensible use. Consequently, the linguistic and informational levels of LLMs' outputs are inherently intertwined, resulting in the user's blended utilization of these functions. In line with Bender et al. (2021, pp. 610–623) and the CASA paradigm (Nass et al., 1994), we agree that we are not sufficiently "account [ing] for the fact that our perception of natural language text, regardless of how it was generated, is mediated by our own linguistic competence and our predisposition to interpret communicative acts as conveying coherent meaning and intent, whether or not they do" (Bender et al., 2021, pp. 610–623). Thus, a key motivation of this research is that it remains under-communicated that the exceptional linguistic proficiency of LLMs does not equate to factual accuracy and real understanding. One attempt to inform users of this discrepancy is by telling them through disclaimers. However, it must be empirically investigated if the sole communication of an AI's fallacies prior to human-AI interaction actually affects users' perceptions of an LLM-generated text.

Thus, the research presented here mainly aimed to examine the effects of disclaimers, that is, information given about strengths and limitations of the AI, on credibility perceptions of a subsequently presented text. Moreover, by using scientific topics as the content of the texts, we considered precisely the lack of factual accuracy in areas where it is essential that people can rely on the provided information. We presented participants with different information on the capabilities of text-generating AI, after which they were asked to evaluate the text and

the AI on various scales. Overall, we found only small differences between the disclaimer types, especially regarding perceived source credibility: In Study 1, the strengths disclaimer induced the highest ratings but did not significantly differ from the limitations disclaimer; in Study 2, the difference between the basic disclaimer and the two limitations disclaimers was significant, but again, the two limitations disclaimers did not differ; for Study 3, the sum contrasts revealed a significant difference between the strengths disclaimer and the other three disclaimer types but it did not significantly differ from the balanced disclaimer, where both strengths and limitations were mentioned. In sum, while it seems to induce a difference in credibility perceptions if the simple definition of a technology is presented compared to if information on its capabilities is shown. However, the data did not support the hypothesis that information about strengths leads to more favorable evaluations and information about pitfalls leads to more negative evaluations. Combined with the overall high credibility ratings and the findings from Study 2 that people hold relatively high expectations toward GenAI's capabilities (machine heuristic), our results can be interpreted as fundamental confidence in an LLMs' output, which remains relatively unaffected by information on the contrary.

The presented research has some limitations, on which we will elaborate in the following. An issue across all three studies was the exclusion of unfit data due to failed manipulation checks, which resulted in uneven group sizes. However, manipulation checks on disclaimer information seem to be a fundamental problem, which other investigations into this manipulation also had to contend with (see Epstein et al., 2022; Metzger et al., 2024). In Study 1, we asked participants to choose between different options regarding the information they were provided with. Those options represented the disclaimer type manipulations between groups. Thus, respondents lacked the comparison and perceived the short definition of AI in the control condition as a representation of its strengths, for instance. In Study 2, participants were asked to indicate only true or not true to a statement that was related to the manipulation they were presented with. Even though this is not a manipulation check in the classical sense, we tried to catch inattentive participants and emphasize which information was presented to them at the same time. However, the uneven group sizes in Study 2 resulted from failed manipulation checks regarding the authorship. A substantial proportion of participants in the co-authorship condition indicated they had read a text written solely by AI. This may have occurred because authorship was not explicitly mentioned before the article presentation and only indicated in the two bylines. In addition, the phrase "checked and revised by a human" could have provided room for interpretation regarding the attribution of authorship. Different degrees of co-authorship and their interpretation by study participants is to be investigated by future studies. In Study 3, we attempted to assess the success of the manipulation by employing a recall test. Participants were presented with two pieces of information from the disclaimer, along with two distractors (true but unmentioned information), for each disclaimer condition. This approach aimed to identify participants who had carefully read the disclaimer and to reinforce the information by presenting it again. Nevertheless, many participants across all conditions (besides the no-disclaimer condition) failed to remember the information and recalled only one correct statement or the distractors. Still, the co-authorship condition suffered less from exclusions due to a change in the co-authorship description. By stating that a human had written the article with the help of an LLM, we assured that the authorship attribution was more explicit and aligned with real-world practices in AI-supported science communication. Moreover, it must be noted that all disclaimer conditions contained a short definition of AI, which might have already given a positive idea of the technologies' capabilities. Since this definition was always presented first, effects such as the primacy effect or the halo effect cannot be ruled out.

Furthermore, it must be considered that for Studies 1 and 3, German-speaking samples were collected, while in Study 2 a US sample was used. This can introduce cultural biases related to language and the perception of technologies. For example, previous research has found intercultural differences regarding attitudes and intentions toward AI use (Ma et al., 2024), but also different approaches and expectations toward AI (Ge et al., 2024). Therefore, cultural differences in the approaches to AI and their interplay with the variables manipulated here cannot be ruled out. Nevertheless, we did not observe strong effects of disclaimer types across all three studies. To be able to show this for two different cultural contexts can also be seen as a strength.

In summary, two major problems must be addressed in future studies: First, research on disclaimer types and prior information must handle manipulation checks in a way that the information is made salient enough to be recognized and memorized appropriately to affect the perception of the subsequent material. What level of knowledge this information is related to and how the respective person weighs it, could depend on context, prior knowledge, and attitudes toward AI and should be investigated simultaneously. In addition, research has shown that cognitive motivation (Buçinca et al., 2021, Liu, 2021) can play a crucial role in XAI's effectiveness. However, it is crucial to consider that high exclusion rates may also reflect a general tendency to disregard disclaimers. Given that only analyzing individuals who could reproduce the disclaimer did not show effects, the overall effectiveness of such disclaimers is questionable. Also, whether experimental conditions that make disclaimers salient enough can be achieved in real-world scenarios must be considered. Moreover, it also raises concerns about whether exclusions based on these criteria introduce systematic bias. To address this issue, robustness analyses should be conducted to ensure that exclusions do not disproportionately impact a critical subset of the sample.

Second, research on co-authorship should consider different levels and manifestations of this authorship option. The manipulation check results in Study 2 suggest that stating a text was written by AI and checked and revised by a human might still be interpreted as full AI authorship. As Bien-Aimé et al. (2024) have shown that people can hold different ideas about the AI contribution to a writing process, querying what respondents understood and imagined by co-authorship can give insights into whom the content and responsibilities are attributed.

Although future research should focus on the effective presentation of preliminary information and meaningful manipulation checks, we would nevertheless like to emphasize the loss of data despite variations of the query about this information. This is a sign that people pay little attention to disclaimers and might also have difficulties interpreting them correctly. The question of how to draw the necessary attention to disclaimers in the first place should therefore be placed before the question of what exact information is communicated. Our research contributes to the impression that a simple byline is not enough. Before considering regulations or restrictions on certain uses, design and placement strategies should be fully explored. The disclaimers used here, despite their brevity, may have created cognitive load and encourage users to skim over them in real-world settings. Additionally, as they were rather general statements, more tailored and contextualized disclaimers are conceivable. For example, warnings can be considered popping up if information is requested that necessitates a data query from reliable sources. Either way, it should be avoided that informative information can be ignored in a similar way to cookies, for instance.

Altogether, the only factor considerably influencing the credibility ratings was a variation of how the article was written, that is, the information presentation in Study 1 (see also Lermann Henestrosa et al., 2023; Tandoc et al., 2020), but less so if or which information was given about possible reliability of the content. Besides, the credibility ratings were rather high across all studies. A plausible explanation for the overall relatively high credibility could be the quality of the articles presented. In line with the heuristic model of persuasion (Chaiken, 2014), easy-to-process cues, such as the length of a text or the number of arguments, could favor minimal information processing and be sufficient to form an impression of a credible article, especially when there are no further reasons for an effortful and systematic analysis. Based on

this, Rader and colleagues (2018, pp. 1–13) found that explanations do not necessarily have an effect on the assessment of the correctness of content but on the awareness of how AI works. This suggests that transparency might not have an immediate influence on the assessment of AI output but perhaps more generally on the perception of AI.

Moreover, since cognitive motivation has been shown to moderate the effectiveness of XAI (Buçinca et al., 2021), and our participants' task was simply to read and evaluate the texts, future studies examining the impact of disclaimers should consider participants' motivation to engage with the articles' content. This could be done by either measuring motivation or, preferably, by manipulating it, for example, by inducing an interaction afterward for which the article's veracity is relevant. Generally, the fact that AI authorship and texts produced by GenAI on nearly any topic are relatively well received by users can be interpreted as a leap of faith in the technologies' capabilities. Nonetheless, given their broad scope of application, it is crucial to develop methods that ensure this foundational confidence does not evolve into a blind trust driven solely by persuasive eloquence but can instead be transformed into informed trust with the necessary skepticism.

The series of studies presented here are among the first to analyze disclaimers on texts produced by GenAI, taking into account the capabilities of LLMs to produce human-like language in a variety of styles, tones, and topics, but always with a convincing confidence that can make it difficult for users to scrutinize the information provided. By presenting different types of information, we addressed key aspects to communicate when people engage with AI-generated content and systematically investigated how the omission of certain pieces of information, for example, might influence perceptions of the text and its author. Moreover, these studies contribute to the emerging field of human-AI-co-authorship by exploring not just the simple comparison of human vs. AI authors but also by considering the more common and realistic scenario of humans and AI collaborating in text creation (Luther et al., 2024). According to a survey from Lermann Henestrosa and Kimmerle (2024), people prefer human authorship across a variety of topics when directly asked for a comparison. However, our findings contribute to the picture that AI authorship is accepted even on more complex scientific topics when readers are presented with it. It remains to be investigated where this rejection in direct comparison stems from and whether doubts and skepticism could be addressed through a human verification authority.

6. Conclusion

Besides the small effects of disclaimer types that point in different directions, this research suggests that prior information about an AI's capabilities hardly affects text and author perceptions. Despite the problems described with the manipulation checks regarding the presented disclaimers, it is questionable how disclaimers in real-world settings can be made that salient to influence interactions taking place at the moment. Future research should work with realistic interactions instead of presenting finished outputs. Moreover, assessing participants' prior attitudes toward LLMs as well as their current level of knowledge, specifically regarding the information highlighted in the disclaimers, could further clarify how disclaimers work depending on different prior knowledge and attitudes. Nonetheless, the present work points against the major effects of information presented before the consumption of AI-generated text on the perceptions of the content or the author. Specifically in the context of scientific information, warnings regarding factual accuracy, for example, might not establish the desired security in dealing with them. This plays the ball back to the companies providing LLMs, which are working on retroactively improving and correcting the models' outputs but need to engage further in their responsibilities to educate and inform users on inappropriate use cases and the limitations of those technologies. Here, we urge developers and applicators to fulfill their duties instead of shifting them to the users and engage in education about the limitations of their products. As our studies suggest, even detailed preliminary information that respondents were forced to read missed to be sufficiently recognized. Therefore, we argue that a footnote cannot fulfill this responsibility adequately, and more intensive efforts should be made to ensure safe use of GenAI.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT 4o mini in order to improve the language of the manuscript and to restructure sentences. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Funding

The research was funded by the Leibniz-Institut fuer Wissensmedien, Tübingen (STB Data Science)

CRediT authorship contribution statement

Angelica Lermann Henestrosa: Writing – original draft, Visualization, Methodology, Formal analysis, Data curation, Conceptualization. **Joachim Kimmerle:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

Nothing to disclose.

Acknowledgments

We would like to thank Emily Herter and Jila Petsch for their valuable contribution to this research, particularly their creation of the material for Studies 1 and 2, as well as their assistance with data collection. Their efforts were crucial to the success of this project.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chbah.2025.100142.

References

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2023). *GPT-4 technical report* (Version 6). *arXiv*. https://doi.org/10.48550/ARXIV.2303.08774

Alufaisan, Y., Marusich, L. R., Bakdash, J. Z., Zhou, Y., & Kantarcioglu, M. (2021). Does explainable artificial intelligence improve human decision-making? *Proceedings of the AAAI Conference on Artificial Intelligence, 35*(8), 6618–6626. https://doi.org/10.1609/aaai.v35i8.16819

Appelman, A., & Sundar, S. S. (2016). Measuring message credibility: Construction and validation of an exclusive scale. *Journalism & Mass Communication Quarterly, 93*(1), 59–79. https://doi.org/10.1177/1077699015606057

Bacon, J., Scheltema, K., & Robinson, B. (2001). Fat phobia scale revisited: The short form. *International Journal of Obesity, 25*(2), 252–257. https://doi.org/10.1038/sj.ijo.0801537

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. *Proceedings of the 2021 CHI conference on human factors in computing systems*. https://doi.org/10.1145/3411764.3445717

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics, 1*(1), 71–81. https://doi.org/10.1007/s12369-008-0001-3

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can Language Models Be too big?. *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. https://doi.org/10.1145/3442188.3445922

Bien-Aimé, S., Wu, M., Appelman, A., & Jia, H. (2024). Who wrote it? News readers' sensemaking of AI/human bylines. *Communication Reports, 1–13*. https://doi.org/10.1080/08934215.2024.2424553

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW1), 1–21. https://doi.org/10.1145/3449287

Bulian, J., Schäfer, M. S., Amini, A., Lam, H., Ciaramita, M., Gaiarin, B., Hübscher, M. C., Buck, C., Mede, N. G., Leippold, M., & Strauß, N. (2023). Assessing Large Language Models on climate information. https://doi.org/10.48550/ARXIV.2310.02932.

Chaiken, S. (2014). The heuristic model of persuasion. In *Social influence (S. 3–39)*. Psychology Press.

Deiana, G., Dettori, M., Arghittu, A., Azara, A., Gabutti, G., & Castiglia, P. (2023). Artificial intelligence and public health: Evaluating ChatGPT responses to vaccination myths and misconceptions. *Vaccines, 11*(7), 1217. https://doi.org/10.3390/vaccines11071217

Diakopoulos, N., & Koliska, M. (2017). Algorithmic transparency in the news media. *Digital Journalism, 5*(7), 809–828. https://doi.org/10.1080/21670811.2016.1208053

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies, 58*(6), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review, 114*(4), 864–886. https://doi.org/10.1037/0033-295X.114.4.864

Epstein, Z., Foppiani, N., Hilgard, S., Sharma, S., Glassman, E., & Rand, D. (2022). Do explanations increase the effectiveness of AI-crowd generated fake news warnings? *Proceedings of the International AAAI Conference on Web and Social Media, 16*, 183–193. https://doi.org/10.1609/icwsm.v16i1.19283

Flanagin, J. A., & Metzger, J. M. (2000). Perceptions of internet information credibility. *Journalism & Mass Communication Quarterly, 77*(3), 515–540. https://doi.org/10.1177/107769900007700304

Fletcher, R., & Nielsen, R. K. (2024). What does the public in six countries think of generative AI in news? *Reuters Institute for the Study of Journalism*. https://doi.org/10.60625/RISJ-4ZB8-CG87

Fox, J. (2016). *Applied regression analysis and generalized linear models* (3rd ed.). Sage.

García-Peñalvo, F., & Vázquez-Ingelmo, A. (2023). What do we mean by GenAI? A systematic mapping of the evolution, trends, and techniques involved in generative AI. *International Journal of Interactive Multimedia and Artificial Intelligence, 8*(4), 7. https://doi.org/10.9781/ijimai.2023.07.006

Ge, X., Xu, C., Misaki, D., Markus, H. R., & Tsai, J. L. (2024). How culture shapes what people want from AI. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3613904.3642660

Greussing, E., Guenther, F., Baram-Tsabari, A., Dabran-Zivan, S., Jonas, E., & Klein-Avraham, I. (2024). Predicting and describing the use of generative AI in science-related information search: Insights from a multinational survey. *Presented at the annual conference of the "science communication" division of the German communication association (DGPuK), Zürich*. https://pure.au.dk/portal/en/publications/predicting-and-describing-the-use-of-generative-ai-in-science-rel.

Hornberger, M., Bewersdorff, A., & Nerdel, C. (2023). What do university students know about artificial intelligence? Development and validation of an AI literacy test. *Computers and Education: Artificial Intelligence, 5*, Article 100165. https://doi.org/10.1016/j.caeai.2023.100165

Hosseini, M., Gao, C. A., Liebovitz, D. M., Carvalho, A. M., Ahmad, F. S., Luo, Y., MacDonald, N., Holmes, K. L., & Kho, A. (2023). An exploratory survey about using ChatGPT in education, healthcare, and research. *PLoS One, 18*(10), Article e0292216. https://doi.org/10.1371/journal.pone.0292216

Jacobs, M., Pradier, M. F., McCoy, T. H., Perlis, R. H., Doshi-Velez, F., & Gajos, K. Z. (2021). How machine-learning recommendations influence clinician treatment selections: The example of antidepressant selection. *Translational Psychiatry, 11*(1), 108. https://doi.org/10.1038/s41398-021-01224-x

Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. *Proceedings of the 28th European conference on information systems (ECIS), an online AIS conference. Proceedings of the 28th European conference on information systems*.

Lee, E.-J. (2024). Minding the source: Toward an integrative theory of human–machine communication. *Human Communication Research, 50*(2), 184–193. https://doi.org/10.1093/hcr/hqad034

Lermann Henestrosa, A., Greving, H., & Kimmerle, J. (2023). Automated journalism: The effects of AI authorship and evaluative information on the perception of a science journalism article. *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2022.107445

Lermann Henestrosa, A., & Kimmerle, J. (2024). Understanding and perception of automated text generation among the public: Two surveys with representative samples in Germany. *Behavioral Sciences, 14*(5), 353. https://doi.org/10.3390/bs14050353

Liu, B. (2021). In AI we trust? Effects of agency locus and transparency on uncertainty reduction in human–AI interaction. *Journal of Computer-Mediated Communication, 26*(6), 384–402. https://doi.org/10.1093/jcmc/zmab013

Lombardi, D., & Sinatra, G. M. (2012). College students' perceptions about the plausibility of human-induced climate change. *Research in Science Education, 42*(2), 201–217. https://doi.org/10.1007/s11165-010-9196-z

Lombardi, D., Sinatra, G. M., & Nussbaum, E. M. (2013). Plausibility reappraisals and shifts in middle school students' climate change conceptions. *Learning and Instruction, 27*, 50–62. https://doi.org/10.1016/j.learninstruc.2013.03.001

Luther, T., Kimmerle, J., & Cress, U. (2024). Teaming up with an AI: Exploring human–AI collaboration in a writing scenario with ChatGPT. AI, 5(3), 1357–1376 https://doi.org/10.3390/ai5030065.

Ma, D., Akram, H., & Chen, H. (2024). Artificial intelligence in higher education: A cross-cultural examination of students' behavioral intentions and attitudes. *International Review of Research in Open and Distance Learning, 25*(3), 134–157. https://doi.org/10.19173/irrodl.v25i3.7703

Metzger, L., Miller, L., Baumann, M., & Kraus, J. (2024). Empowering calibrated (Dis-) Trust in conversational agents: A user study on the persuasive power of limitation disclaimers vs. Authoritative style. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–19. https://doi.org/10.1145/3613904.3642122

Molina, M. D., & Sundar, S. S. (2024). Does distrust in humans predict greater trust in AI? Role of individual differences in user responses to content moderation. *New Media & Society, 26*(6), 3638–3656. https://doi.org/10.1177/14614448221103534

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*(1), 81–103. https://doi.org/10.1111/0022-4537.00153

Nass, C., Steuer, J., & Tauber, E. R. (1994). *Computers are Social Actors*, 72–78. https://doi.org/10.1145/191666.191703

Palmeira, M., & Spassova, G. (2015). Consumer reactions to professionals who use decision aids. *European Journal of Marketing, 49*(3/4), 302–326. https://doi.org/10.1108/EJM-07-2013-0390

Papenmeier, A., Englebienne, G., & Seifert, C. (2019). How model accuracy and explanation fidelity influence user trust. *arXiv*. https://doi.org/10.48550/ARXIV.1907.12652 Version 1.

Phillips, P. J., Hahn, C. A., Fontana, P. C., Yates, A. N., Greene, K., Broniatowski, D. A., & Przybocki, M. A. (2021). *Four principles of explainable artificial intelligence (No. NIST IR 8312; S. NIST IR 8312)*. U.S.: National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8312

Rader, E., Cotter, K., & Cho, J. (2018). Explanations as mechanisms for supporting algorithmic transparency. *Proceedings of the 2018 CHI conference on human factors in computing systems*. https://doi.org/10.1145/3173574.3173677

Said, N., Potinteu, A. E., Brich, I. R., Buder, J., Schumm, H., & Huff, M. (2022). An artificial intelligence perspective: How knowledge and confidence shape risk and opportunity perception. *PsyArXiv*. https://doi.org/10.31234/osf.io/5zvha

Seo, H., Xiong, A., & Lee, D. (2019). Trust it or not: Effects of machine-learning warnings in helping individuals mitigate misinformation. *Proceedings of the 10th ACM conference on web science* (pp. 265–274). https://doi.org/10.1145/3292522.3326012

Shahsavar, Y., & Choudhury, A. (2023). User intentions to use ChatGPT for self-diagnosis and health-related purposes: Cross-sectional survey study. *JMIR Human Factors, 10*, Article e47564. https://doi.org/10.2196/47564

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies, 146*, Article 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

Shin, D., & Park, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior, 98*, 277–284. https://doi.org/10.1016/j.chb.2019.04.019

Sulmont, E., Patitsas, E., & Cooperstock, J. R. (2019). Can you teach me to machine learn?. *Proceedings of the 50th ACM technical symposium on computer science education* (pp. 948–954). https://doi.org/10.1145/3287324.3287392

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

Sundar, S. S. (2008). The main model: A heuristic approach to understanding technology effects on credibility. In J. M. Metzger, & J. A. Flanagin (Eds.), *(Hrsg.), digital media, youth, and credibility (S. 73–100)*. The MIT Press.

Sundar, S. S., & Kim, J. (2019). Machine heuristic: When we trust computers more than humans with our personal information. *Proceedings of the 2019 CHI conference on human factors in computing systems*. https://doi.org/10.1145/3290605.3300768

Tandoc, E. C., Yao, L. J., & Wu, S. (2020). Man vs. Machine? The impact of algorithm authorship on news credibility. *Digital Journalism, 8*(4), 548–562. https://doi.org/10.1080/21670811.2020.1762102

Tully, S., Longoni, C., & Appel, G. (2025). Express: Lower artificial intelligence literacy predicts greater AI receptivity. *Journal of Marketing*. , Article 00222429251314491. https://doi.org/10.1177/00222429251314491

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly, 27*(3), 425–478. https://doi.org/10.2307/30036540

Vered, M., Livni, T., Howe, P. D. L., Miller, T., & Sonenberg, L. (2023). The effects of explanations on automation bias. *Artificial Intelligence, 322*, Article 103952. https://doi.org/10.1016/j.artint.2023.103952

Wang, S. (2021). Moderating uncivil user comments by humans or machines? The effects of moderation agent on perceptions of bias and credibility in news content. *Digital Journalism, 9*(1), 64–83. https://doi.org/10.1080/21670811.2020.1851279

Wang, S., & Huang, G. (2024). The impact of machine authorship on news audience perceptions: A meta-analysis of experimental studies. *Communication Research, 51*(7), 815–842. https://doi.org/10.1177/00936502241229794

Wei, J., Karina, N., Chung, H. W., Jiao, Y. J., Papay, S., Glaese, A., Schulman, J., & Fedus, W. (2024). Measuring short-form factuality in large language models. *arXiv*. https://doi.org/10.48550/ARXIV.2411.04368 Version 1.

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., & Zhu, J. (2019). Explainable AI: A brief survey on history, research areas, approaches and challenges. In J. Tang, M.-Y. Kan, D. Zhao, S. Li, & H. ZanHrsg (Eds.), *Natural Language Processing and Chinese computing (Bd. 11839, S* (pp. 563–574). Springer International Publishing. https://doi.org/10.1007/978-3-030-32236-6_51.

C Eidesstattliche Erklärung

Ich erkläre hiermit, dass ich die zur Promotion eingereichte Arbeit selbständig verfasst, nur die angegebenen Quellen und Hilfsmittel benutzt und wörtlich oder inhaltlich übernommene Stellen als solche gekennzeichnet habe. Ich erkläre, dass die Leitlinien zur Sicherung guter wissenschaftlicher Praxis der Universität Tübingen (Beschluss des Senats vom 11.2.2021) beachtet wurden. Ich versichere an Eides statt, dass diese Angaben wahr sind und dass ich nichts verschwiegen habe. Mir ist bekannt, dass die falsche Abgabe einer Versicherung an Eides statt mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft wird.

Datum Unterschrift

D Erweiterte Erklärung zur Verwendung generativer Künstlicher Intelligenz

Mir ist bewusst, dass die Nutzung mittels generativer KI erstellter Texte oder Inhalte keine Garantie für deren Qualität gewährleistet und ich die Verantwortung trage, falls es durch die Verwendung solcher Hilfsmittel zu fehlerhaften Inhalten, zu Verstößen gegen das Datenschutzrecht, Urheberrecht oder zu wissenschaftlichem Fehlverhalten (z.B. Plagiate) kommt.

Ich versichere außerdem,

- dass ich mich generativer KI-Tools lediglich als Hilfsmittel bedient habe und in der vorliegenden Arbeit mein gestalterischer Einfluss überwiegt,
- dass ich angegeben habe, welche generativen KI-Tools ich zu welchem Zweck und in welchem Umfang eingesetzt habe:
  - Ich habe generative KI und KI-basierte Anwendungen (GPT-4o, Copilot 4.0, DeepL, Grammarly) zum Zwecke der Übersetzung, Verbesserung der Lesbarkeit sowie Formulierung dieser Dissertation verwendet. Zudem habe ich mittels Copilot eine Lay Summary erstellt. Nach der Nutzung dieser Dienste habe ich den Inhalt nach Bedarf überprüft und bearbeitet und übernehme die volle Verantwortung für den Inhalt der veröffentlichten Dissertation.
  - Ich habe generative KI (GPT-3, GPT-4) zur Materialerstellung verwendet und dieses in den jeweiligen Publikationen gekennzeichnet.

Die vorliegende Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt. Mir ist bekannt, dass ein Verstoß gegen die genannten Punkte prüfungsrechtliche Konsequenzen haben und insbesondere dazu führen kann, dass die Prüfungsleistung mit „nicht ausreichend“ bzw. die Studienleistung mit „nicht bestanden“ bewertet wird und bei mehrfachem oder schwerwiegendem Täuschungsversuch eine Exmatrikulation erfolgen bzw. ein Verfahren zur Entziehung eines eventuell verliehenen akademischen Titels eingeleitet werden kann.

Datum Unterschrift