Striking the Balance in Using LLMs for Fact-Checking: A Narrative Literature Review

Conference Paper · August 2024

DOI: 10.1007/978-3-031-71210-4_1

CITATIONS
14

READS
631

4 authors:

Laurence Dierickx
Université Libre de Bruxelles
34 PUBLICATIONS **151** CITATIONS
SEE PROFILE

Arjen van Dalen
University of Southern Denmark
74 PUBLICATIONS **2,688** CITATIONS
SEE PROFILE

Andreas Lothe Opdahl
University of Bergen
173 PUBLICATIONS **4,564** CITATIONS
SEE PROFILE

Carl-Gustav Linden
University of Bergen
76 PUBLICATIONS **918** CITATIONS
SEE PROFILE

Striking the Balance in Using LLMs for Fact-Checking: A Narrative Literature Review

Laurence Dierickx[1,2(✉)], Arjen van Dalen[2], Andreas L. Opdahl[1], and Carl-Gustav Lindén[1]

[1] Department of Information Science and Media Studies, University of Bergen, Bergen, Norway
Laurence.Dierickx@uib.no

[2] Digital Democracy Centre, University of Southern Denmark, Odense, Denmark

Abstract. The launch of ChatGPT at the end of November 2022 triggered a general reflection on its benefits for supporting fact-checking workflows and practices. Between the excitement of the availability of AI systems that no longer require the mastery of programming skills and the exploration of a new field of experimentation, academics and professionals foresaw the benefits of such technology. Critics have raised concerns about the fairness and of the data used to train Large Language Models (LLMs), including the risk of artificial hallucinations and the proliferation of machine-generated content that could spread misinformation. As LLMs pose ethical challenges, how can professional fact-checking mitigate risks? This narrative literature review explores the current state of LLMs in the context of fact-checking practice, highlighting three key complementary mitigation strategies related to education, ethics and professional practice.

Keywords: Fact-checking · Large Language Models · Risk mitigation

1 Introduction

The launch of ChatGPT on 30 November 2022 marked a significant milestone in the integration of artificial intelligence (AI) into newsrooms [1]. Its user-friendly interface and the elimination of the need for extensive programming skills accelerated this process [2]. Building on the success of the Generative Pre-trained Transformer (GPT) architecture, Large Language Models (LLMs) such as ChatGPT use machine learning on large databases to generate content [3,4]. As foundational models, LLMs operate by learning patterns and information from vast datasets, which presents complexities in ensuring ethical use and reliable results in fact-checking, although they are seen as an opportunity to improve workflows and augment professional practice [5–7].

Ethical concerns arise from the opacity of data collection and training datasets, where biased and inaccurate data can skew results [8,9]. For instance,

M. Preuss et al. (Eds.): MISDOOM 2024, LNCS 15175, pp. 1–15, 2024.
https://doi.org/10.1007/978-3-031-71210-4_1

studies have shown that ChatGPT's primary sources are not immune to bias, error or partisanship [10–12]. Its use of copyrighted data increases the risk of plagiarism, as the system tends to reproduce sentences from its training dataset [13]. Another significant challenge lies in its potential to spread misinformation and disinformation [14,15]. Because LLMs generate content that is not easily distinguishable from human-generated information, they are potentially harmful tools [16]. However, the generation of misleading content is not always intentional and also includes artificial hallucinations, a well-known problem, which refers to the generation of fact-like claims that contradict real-world facts [17,18]. Further, the black-box nature of LLMs raises questions of trustworthiness and accountability, exacerbating existing concerns about AI and generative artificial intelligence (GAI) [19–22].

Given the ethical challenges and limitations of these technologies, what is the potential for professional fact-checking, and how can the risks be mitigated? This paper addresses this question through a narrative literature review; a qualitative research tool used strategically to allow flexibility in exploring different methodologies. Its primary aim is to provide a comprehensive overview of existing knowledge in an emerging field [23–25]. This research builds on previous related research, which consist of a systematic review of automated fact-checking from an end-user perspective [26] and a study on dealing with factuality in LLMs [27]. The paper collection is essentially based on a snowball method, a technique in which initial sources lead the researcher to further relevant studies. This iterative and qualitative approach facilitated the exploration of recent advances in LLMs in journalism studies and information science. In this research, fact-checking is approached as a specialised sub-genre of journalism that fulfils the promise of objective and responsible reporting [28,29].

To provide an in-depth understanding of the topic, this paper is divided into two main sections: (1) outlining the contours of professional fact-checking by describing its processes and exploring its intersection with technology, and (2) examining the potential applications of LLMs in fact-checking and formulating strategies for effectively mitigating the risks. To highlight the implications associated with LLMs in fact-checking, this study adopts an interdisciplinary approach [30]. We identified three complementary strategies: education through AI literacy, ethics through promoting human oversight to ensure accountability, and professional practice through improving prompt engineering –the instructions provided by the user to the system– to achieve better results.

2 Fact-Checkers and Fact-Checking Technology

In journalism, verification is an ethical standard ensuring the reliability, accuracy, credibility and truthfulness of the news [31–33]. Verification takes place before publication, whereas fact-checking takes place after publication. It has evolved as a sub-genre in journalism, following the need to verify political discourses and the content propagated on social media [34,35]. In practice, fact-checking is an iterative and time-consuming process that can be standardised

into five main stages. These stages are well documented in the literature, particularly in discussions of automated fact-checking techniques [36,37].

1. **Media monitoring:** This stage involves actively monitoring various social media platforms and political sources to identify claims or statements needing verification or worth checking [38]. Monitoring tasks can be supported by AI technology, as it is time-consuming when performed manually [39].
2. **Selecting the claim to check:** Fact-checkers decide which claims or information to check based on relevance, potential impact and other criteria. This stage primarily requires critical thinking skills and mandates that fact-checkers also question their possible selection bias [40,41]. Machine learning and natural language applications can be used in this process, for claim detection or claim selection, including assessing the newsworthiness of the claim [42–44].
3. **Verification and evidence gathering:** Verification aims to assess the truth, accuracy or validity of a claim [33]. For this, fact-checkers thoroughly investigate selected claims, using traditional journalistic methods such as expert interviews, digital tools and open-source information to gather evidence. The techniques and methods used depend on the nature of the claim being verified, whether it is textual or audiovisual. Verification of social media content is particularly challenging as it involves the assessment of user-generated content, which has greater potential for manipulation, alteration and removal from context [35]. This stage of the process can be supported by a wide range of digital tools, of which the AI-based ones are mostly concerned with assigning credibility ratings to information [35,45].
4. **Giving a verdict:** Based on the evidence collected, fact-checkers determine the accuracy of the claim and assign a verdict, such as "True", "False" or "Misleading". The rating used may differ from one fact-check to another, and may also differ when the same claim is fact-checked by two different organisations [46]. Recognising the possibility of nuances between truth and falsehood, some fact-checking organisations have developed a rating scale, but this can be difficult to use at the operational level [47]. Machine learning predictive algorithms are likely to support this stage [42,44,48].
5. **Producing the narrative:** The final stage involves articulating the findings and evidence comprehensively and communicating them to audiences through articles, reports or other content. Transparent storytelling is rooted in open-source practices and involves the integration of digital skills and visual evidence collected during the investigation. As such, fact-checkers not only present facts but also reveal the methods used in fact-checking [49]. This stage could be assisted by natural language generation methods, but we did not find any evidence of effective use in the literature review.

2.1 From Traditional to Technological Skills

Technological advances aim to support fact-checking practices in a process that is acknowledged as being time-consuming. They unfold in a context where journalists have always had an ambivalent relationship with technology. While they are

open to technologies that benefit their work, they balance technological opportunities and challenges with their professional autonomy and identity [50–52].

The adoption of technologies in journalism is influenced by a multifaceted interplay of socio-professional factors, including professional backgrounds, individual skills, and organisational contexts [53]. A recent study in Norway revealed a nuanced integration of fact-checking into journalism, highlighting the tension between preserving traditional skills and embracing innovative methods. Notably, fact-checkers in the study relied increasingly on digital verification tools [54]. Swedish newsrooms face analogous challenges in navigating organisational dynamics, time constraints, and barriers to skills acquisition [55].

Verifying the accuracy of sources and content is particularly complex in the dynamic landscape of social media platforms, where user-generated content often serves as a conduit for misinformation and disinformation [35]. The complexity escalates with the potential for misrepresentation when critical context is omitted [56]. The application of open source intelligence (OSINT) methods and tools, which leverage publicly available resources such as geolocation data, facial recognition technology, and web archives, has significant potential to revolutionize investigative techniques by enhancing accuracy and depth of analysis through the integration of digital tools [49,57,58]. Despite their potential benefits, fact-checkers have not yet widely adopted OSINT methods due to a lack of awareness and skills in using such tools [40]. More generally, fact-checkers often rely on familiar tools for research and verification, despite the plethora of digital tools at their disposal [53,59].

Developing a critical mindset coupled with logical reasoning remains one of the fundamental skills for fact-checkers [53], distinguishing human judgement from automated systems [60]. Emotional and social intelligence, essential for nuanced journalistic judgement, cannot be replicated in computer code; AI-based systems may mimic human patterns but do not possess human understanding [61]. Therefore, integrating technological innovation in journalism and fact-checking is a complex issue, influenced more by social and cultural considerations than mere technical advances. Furthermore, concerns remain about the potential of technology to compromise professional autonomy [62], underscoring the need for preserving human agency and professional values [52,63,64].

2.2 Practices of Transparency vs. Opacity of Technology

AI-based technologies challenge ethical fact-checking practices in a number of ways. The majority of European fact-checking organisations typically adhere to a newsroom model, where teams operate under journalistic standards centred on truthfulness, fairness and accuracy [65,66]. Their professional identity also depends on a strong commitment to transparency, particularly among those who belong to the International Fact-Checking Network (IFCN) or the European Fact-Checking Standards Networks (EFSCN). Membership of these networks mandates transparency about sources, funding, methods and correction policies, thereby aligning journalistic ethics with accountability standards [53]. However,

not all fact-checkers prioritise transparency as their primary professional norm, privileging accuracy, impartiality, objectivity or independence instead [28].

Fact-checkers whose organisations are signatories to the IFCN standards are granted access to Facebook's third-party fact-checking programme, consisting of a tool that facilitates the identification of checkworthy viral content. The reliance of Facebook's algorithm on an opaque process and lack of accountability for its behaviour does not prevent its use, although fact-checkers have criticised its opacity [53]. This is usually the case when fact-checkers use technology: it is not common for systems to be documented for their end users, and fact-checkers often lack awareness of the technological intricacies that underpin their functionality [67]. This highlights the critical role of explainability in AI-based technologies, which aims to increase trust and reliability by understanding the reasoning behind outcomes [10]. In AI ethics, explainability complements transparency as it also helps to assign responsibility for unintended consequences [68]. While explainability does not ensure trustworthy AI, it is a critical tool for fostering trust in AI [69,70]. This need for accountability is particularly evident in automated fact-checking technology, where concerns about trust and algorithmic accountability remain significant barriers to widespread adoption [71].

The limited adoption of computational tools by fact-checkers is not only due to transparency issues, but also to their specialisation in specific fact-checking functions, requiring manual effort and lacking seamless integration [72]. Despite advances in AI-based solutions, skepticism remains among fact-checkers, especially regarding critical processes such as contextualisation and the management of real-time constraints [67,72]. Furthermore, fact-checking technologies developed outside the field often diverge from the expertise and values of fact-checkers [26,48], whereas human-in-the-loop approaches have shown significant benefits [73,74]. This underlines that technology alone cannot adequately address the complex societal challenges posed by information disorders.

3 Use Cases for LLMs

The integration of LLMs into fact-checking practices is based on the premise that humans may struggle to manage vast amounts of disinformation [75]. In this narrative literature review, we do not distinguish between specific systems, as ChatGPT is not the only LLM available on the market. Instead, it provides a comprehensive overview of potential risks and mitigation strategies based on the current state of knowledge.

3.1 Possibilities and Identification of Risk Factors

Research has identified several ways in which AI-based technology is likely to improve fact-checking workflows and practices, recognising that verification also relies on different standards, individual judgements, first-hand experience and relationships with sources [76]. In investigative journalism, AI methods have been used to uncover hidden patterns in large amounts of data, although these

efforts have produced modest results due to unique story requirements and issues of data availability and quality [77]. AI methods in fact-checking can be used for social media monitoring, assessing the newsworthiness of a claim, stance detection, searching for previous fact-checks, audiovisual content verification, semi-automated classification, feature extraction, writing tasks or disseminating fact-checks [72,78]. Similarly, LLMs are considered to have potential for extracting data from large amounts of information, providing contextual information, detecting disinformation, analysing data, summarising text, detecting stance, writing and supporting multilingual tasks [4,30].

In their pioneering study conducted in Spain, Cuartielles et al. identified several benefits for using LLMs in fact-checking, such as the ease of data collection and the ability of AI tools such as ChatGPT to provide synthetic and rapid information, and contextual insights for fact-checking processes [6]. The study also highlighted the potential for self-learning, improving the performance of the system and contributing to the development of more specialised tools devoted to specific tasks [6]. Practically, LLMs are likely to support each stage of the manual fact-checking process as illustrated in Table 1.

Table 1. Identifying potential for using LLMs to support the fact-checking process

Fact-Checking Stage	LLMs Associated Tasks
1. Media monitoring	Extracting data from multiple sources, including social media corpora, to create rich datasets for subsequent analysis. Processing and organising large amounts of data.
2. Selecting the claim to check	Identifying and prioritising claims for review, including providing contextual analysis to support the decision.
3. Verification and evidence gathering	Cross-referencing the identified claims with sources or databases of factual information. Preparing summaries of the information gathered. Analysing the collected data, assessing the accuracy based on the analysis.
4. Giving a verdict	Providing contextual understanding to support judgment.
5. Producing the narrative	Generating texts based on pre-defined templates. Supporting the production of clear and structured narratives.

Despite all these promises, LLMs remain untrustworthy for most practical fact-checking uses due to several current challenges, including dealing with artificial hallucinations [10,79]. Unlike intrinsic hallucinations, which contradict the content of the source, extrinsic hallucinations in a LLM cannot be verified by the source because the systems do not provide the source on which they base their results [80]. Hallucinations are a well-known phenomenon and they are explained by the fact that the system has been trained to generate maximally plausible sequences of words, not to internalise world knowledge, although research also pointed out explanations related to the use of large amounts of unsupervised data, a lack of quality in the training data and the black-box nature of the system [18,81,82]. In addition, LLMs face limitations in contextual understanding and may lack access to real-time or updated data [4,18]. Retrieval-Augmented Generation (RAG) is a technique used to combat hallucinations. It consists of augmenting the system with external data, assuming that contextual information from external sources will improve text generation [2,83]. However, this technical solution is inaccessible to fact-checkers, who can enhance their prompting techniques by providing additional information.

When dealing with long documents, summarising can lead to important context being omitted [3]. Further, there is a gap in understanding the origin of the source, with a noticeable lack of source transparency in the results generated by the system [2,6,13,68,79]. Moreover, bias in the collection and training of data remains an open issue [4,79]. Using LLMs in fact-checking also introduces several risks that demand careful considerations. These risks encompass the potential creation and dissemination of unintentional disinformation (artificial hallucinations) or intentional misinformation that is difficult to detect [10,13,84]. The confident tone of the output can also be misleading, especially since the system might misinterpret the data and have a limited understanding of technical terms [2,79]. Further, the articulate communication of LLMs can encourage a misleading tendency towards anthropomorphism-the attribution of human characteristics to the system [2]. The use of LLMs in fact-checking poses other challenges related to artificial hallucinations, which are problematic for generating expert knowledge [18], and to copyrighted data, involving a risk of plagiarism [10,13]. Another concern is the phenomenon of deskilling, already observed during previous industrial revolutions [85]. For fact-checkers, it refers to the decline in critical thinking and problem-solving skills [4,86].

The hype surrounding LLMs may have social consequences, in terms of anxiety about workforce displacements [13,87]. Further, the perceived credibility and reliability of LLMs is likely to undermine the user's trust in technology, which plays a pivotal role in shaping interactions between humans and AI systems [88–90]. Hence, fact-checkers may see LLMs as a supplementary tools rather than a means to augment human possibilities [6]. Research in this area is still nascent due to the novelty of the technology, but these early indicators show the ambivalent nature of LLMs and highlight the need for risk mitigation, to create the conditions for trustworthy fact-checking systems, a prerequisite for uses in journalism and fact-checking [91].

3.2 Strategies for Risk Mitigation

Three key strategies for mitigating the risks associated with LLMs are explored in this section: promoting AI literacy, encouraging human oversight and collaboration, and working on effective prompting techniques to combat artificial hallucinations. The focus is on formulating a comprehensive set of strategies that address the challenges and limitations of the technology, all within the context of fostering responsible and ethical use of LLMs in fact-checking.

Promoting AI Literacy. Fostering a robust understanding of AI among journalists and fact-checkers has never been more important than in the age of generative AI technology. Deuze and Beckett define AI literacy as "the knowledge and beliefs about artificial intelligence that facilitate its recognition, management and application" [92, p. 1913]. It is, therefore, not just about applications and practices but also has a strong ethical dimension. Further, a lack of AI literacy could have negative consequences [13]. Initial or ongoing training programmes can help to understand these challenges, equip professionals with the appropriate skills, develop a critical mindset to prevent further risks and prevent being fooled by the convincing tone of LLMs [2,75,93]. Clear guidelines are also needed in newsrooms to guide responsible use and best ethical practices, including the scope of LLMs and other GAI applications. In this context, it is noteworthy that until the launch of ChatGPT, the ethical use of AI was rather overlooked by self-regulatory bodies in Europe and news media organisations. The situation has changed rapidly with the development of LLMs, prompting a great deal of reflection that has led several newsrooms to adopt specific guidelines to manage the limitations and mitigate the risks associated more broadly with AI-based systems. In addition, two European press councils - in the UK (Impress) and Belgium (Raad voor de Journalistiek) - have already adapted their codes of journalistic ethics with an emphasis on accuracy and transparency [94].

Human Oversight and Collaboration. Maintaining human oversight to balance accuracy when there is no guarantee of algorithmic accountability is a strategy to mitigate potential adverse effects [18]. It enables fact-checkers to use LLMs for close supervision [6]. It is, therefore, also a mitigation strategy that preserves the autonomy and authority of fact-checkers, or in other words, their role in evaluating and validating claims [41]. The principle of human oversight involves collaboration with humans, according to a human-in-the-loop approach that incorporates external knowledge, tools and multimodal information augmentation. Such a collaborative approach improves accuracy, reinforces the ethical use of LLMs in journalism and fact-checking, and improves the overall user experience [95]. It is also deemed to get better results from the perspective of a human-machine association for human augmentation [10]. The majority of newsrooms that have adopted ethical guidelines for using AI and GAI have placed this principle at the forefront of their recommendations, recognising the human responsibility [94]. For fact-checkers, it can be considered a means to maintain a high level of transparency that counterbalances the opacity of the system.

Developing Prompting Techniques for Combating Hallucinations. The results produced by LLMs are often inexplicable and can be prone to artificial hallucinations. They can also give different answers to the same question [2]. Research has shown that the development and implementation of creative prompting techniques can significantly improve the explainability of these tools [75] and effectively mitigate the generation of fabricated content and artificial hallucinations [18,96]. Prompting involves programming the system using natural language, making it accessible to users without requiring extensive computer skills. It is seen as a valuable tool for improving the quality of results and can potentially revolutionise problem-solving in various domains [99]. To improve prompting, experimenting with different instructions can help to assess the quality of the outcome [96]. Incorporating specific context and expectations into the prompt can also improve the accuracy of the result [97,98]. In addition, prompting can be used to generate new prompts, allowing for self-adaptation and the potential for continuous improvement [99]. Prompting has also proven effective in detecting misinformation and disinformation, leveraging the world knowledge and inferential abilities of LLMs [30].

4 Discussion and Conclusion

This narrative literature review explored the ambivalent nature of LLMs and assessed their potential use in fact-checking, along with considering their limitations and the risks they may pose to the quality of information output. The main findings highlight the need to develop (G)AI literacy programmes tailored for fact-checkers, the importance of ensuring ethical use through a human-in-the-loop approach and the benefits of researching and developing prompting strategies to improve the quality of results, especially by addressing hallucination-related challenges.

Concerning automated fact-checking (AFC), LLMs may play a critical role in facilitating explainability. Although it does not guarantee that any AI-based system will be flawless, explainability can be seen as an instrument for fostering trust in AFC systems by providing a level of transparency consistent with the professional standards of fact-checkers and addressing criticisms about the opacity of systems [69]. Another benefit of explainability is that it aims to provide a comprehensive human-level understanding of how systems work in processes and decision-making [100]. However, achieving explanations that accurately reflect the inner workings of LLMs remains a challenging and ongoing area of research. Moreover, LLMs may never be able to explain how they work with an acceptable level of accuracy, fairness and reliability, mainly due to the question of the sources used in the generated results and the challenges of dealing with artificial hallucinations. When asked to explain their behaviour, LLMs may hallucinate explanations, forcing them to generate plausible-sounding but potentially inaccurate answers. This paper does not discuss tools to detect AI-generated text due to their inherent limitations regarding accuracy and reliability [101]. Relying solely on these tools could lead to overlooking nuanced aspects of fact-checking,

such as context, intent and veracity of information. Therefore, an emphasis on content quality ensures a more holistic approach, combining the strengths of LLMs with human oversight to maintain professional standards. In addition, from a professional perspective, verifying the truthfulness of the content is more important than knowing how it was produced, especially since LLMs can be used ambivalently to inform and disinform. Further, this paper provides a comprehensive analysis of the state of the art, identifying key challenges and promising solutions that require further investigation. It paves the way for future empirical research to assess the effectiveness of these mitigation strategies. It also encourages a thorough assessment of the benefits and risks of using LLMs for fact-checking and, more broadly, for journalism, recognising that neither transparency nor accountability is sufficient to meet the ethical requirement of respect for facts.

Acknowledgments. This research was funded by EU CEF Grant No. 101158604.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Beckett, C., Yaseen, M.: Generating Change. A global survey of what news organisations are doing with AI (2023). https://static1.squarespace.com/static/64d60527c01ae7106f2646e9. Accessed 22 June 2024
2. Augenstein, I., et al.: Factuality challenges in the era of large language models. arXiv preprint arXiv:2310.05189 (2023)
3. Aydin, Ö., Karaarslan, E.: Is ChatGPT leading generative AI? What is beyond expectations? Acad. Platform J. Eng. Smart Syst. **11**(3), 118–134 (2023)
4. Yenduri, G., et al.: GPT (Generative pre-trained transformer)-a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. arXiv, Eprint: arXiv:2305.10435v2 (2023)
5. Unver, H.A.: Emerging Technologies and Automated Fact-Checking: Tools, Techniques and Algorithms (2023). SSRN: https://ssrn.com/abstract=4555022
6. Cuartielles, R., Ramon-Vegas, X., Pont-Sorribes, C.: Retraining fact-checkers: the emergence of ChatGPT in information verification. Profesional de la información/Inf. Profess. **32**(5) (2023)
7. Wolfe, R., Mitra, T.: The impact and opportunities of generative AI in fact-checking. In: The 2024 ACM Conference on Fairness, Accountability, and Transparency, pp. 1531–1543 (2024)
8. Budach, L., et al.: The effects of data quality on machine learning performance. arXiv preprint arXiv:2207.14529 (2022)
9. Gudivada, V.N., Apon, A., Ding, J.: Data quality considerations for big data and machine learning: going beyond data cleaning and transformations. Int. J. Adv. Softw. **10**(1), 1–20 (2017)
10. Dwivedi, Y.K., et al.: So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. Int. J. Inf. Manage. **71**, 102642 (2023)

11. Hartmann, J., Schwenzow, J., Witte, M.: The political ideology of conversational AI: converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. arXiv preprint arXiv:2301.01768 (2023)
12. Fujimoto, S., Takemoto, K.: Revisiting the political biases of ChatGPT. Front. Artif. Intell. **6**, 1232003 (2023)
13. Jones, B., Luger, E., Jones, R.: Generative AI & Journalism: A Rapid Risk-Based Review. Edinburgh Research Explorer, University of Edinburgh (2023)
14. Wach, K., et al.: The dark side of generative artificial intelligence: a critical analysis of controversies and risks of ChatGPT. Entrep. Bus. Econ. Rev. **11**(2), 7–30 (2023)
15. Bontcheva, K., et al.: Generative AI and Disinformation: Recent Advances, Challenges, and Opportunities. European Digital Media Observatory (2024)
16. Spitale, G., Biller-Andorno, N., Germani, F.: AI model GPT-3 (dis) informs us better than humans. arXiv preprint arXiv:2301.11924 (2023)
17. Hanley, H.W.A., Durumeric, Z.: Machine-made media: monitoring the mobilization of machine-generated articles on misinformation and mainstream news websites. arXiv preprint arXiv:2305.09820 (2023)
18. Rawte, V., Sheth, A., Das, A.: A survey of hallucination in large foundation models. arXiv preprint arXiv:2309.05922 (2023)
19. Dignum, V.: Responsible artificial intelligence: designing AI for human values. ITU J. ICT Discov. **1**, 1–8 (2017)
20. Johnson, B., Smith, J.: Towards ethical data-driven software: filling the gaps in ethics research & practice. In: 2021 IEEE/ACM 2nd International Workshop on Ethics in Software Engineering Research and Practice (SEthics) (2021)
21. Khan, A.A., et al.: Ethics of AI: a systematic literature review of principles and challenges. In: Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering, pp. 383–392 (2022)
22. Stahl, B.C.: Concepts of ethics and their application to AI. In: Artificial Intelligence for a Better Future. SRIG, pp. 19–33. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-69978-9_3
23. Jahan, N., Naveed, S., Zeshan, M., Tahir, M.A.: How to conduct a systematic review: a narrative literature review. Cureus **8**(11), e864 (2016)
24. Baumeister, R.F., Leary, M.R.: Writing narrative literature reviews. Rev. Gen. Psychol. **1**(3), 311–320 (1997)
25. Rother, E.T.: Systematic literature review X narrative review. Acta Paulista de Enfermagem **20**, v–vi (2007)
26. Dierickx, L., Lindén, C.-G., Opdahl, A.L.: Automated fact-checking to support professional practices: systematic literature review and meta-analysis. Int. J. Commun. **17**, 21 (2023)
27. Dierickx, L., Lindén, C., Opdahl, A.: The information disorder level (IDL) index: a human-based metric to assess the factuality of machine-generated content. In: Multidisciplinary International Symposium On Disinformation in Open Online Media, pp. 60–71 (2023)
28. Singer, J.B.: Border patrol: the rise and role of fact-checkers and their challenge to journalists' normative boundaries. Journalism **22**(8), 1929–1946 (2021)
29. Mena, P.: Principles and boundaries of fact-checking: journalists' perceptions. Journal. Pract. **13**(6), 657–672 (2019)
30. Chen, C., Shu, K.: Combating misinformation in the age of LLMs: opportunities and challenges. arXiv preprint arXiv:2311.05656 (2023)

31. Shapiro, I., Brin, C., Bédard-Brûlé, I., Mychajlowycz, K.: Verification as a strategic ritual: how journalists retrospectively describe processes for ensuring accuracy. Journal. Pract. **7**(6), 657–673 (2013)
32. Martin, N., Comm, B. A.: Information verification in the age of digital journalism. In: Special Libraries Association Annual Conference, Vancouver (2014)
33. Hermida, A.: Tweets and truth: journalism as a discipline of collaborative verification. Journal. Pract. **6**(5–6), 659–668 (2012)
34. Graves, L., Amazeen, M.A.: Fact-checking as idea and practice in journalism. In: Oxford Research Encyclopedia of Communication. Oxford University Press, Oxford (2019)
35. Brandtzaeg, P.B., Lüders, M., Spangenberg, J., Rath-Wiggins, L., Følstad, A.: Emerging journalistic verification practices concerning social media. Journal. Pract. **10**(3), 323–342 (2016)
36. Konstantinovskiy, L., Price, O., Babakar, M., Zubiaga, A.: Toward automated fact-checking: developing an annotation schema and benchmark for consistent automated claim detection. Digit. Threats Res. Pract. **2**(2), 1–16 (2021)
37. Vlachos, A., Riedel, S.: Fact checking: task definition and dataset construction. In: Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, pp. 18–22 (2014)
38. Sheikhi, G., Touileb, S., Khan, S.: Automated claim detection for fact-checking: a case study using Norwegian pre-trained language models. In: Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa), pp. 1–9 (2023)
39. Al-Ghamdi, L.M.: Towards adopting AI techniques for monitoring social media activities. Sustain. Eng. Innov. **3**(1), 15–22 (2021)
40. Himma-Kadakas, M., Ojamets, I.: Debunking false information: investigating journalists' fact-checking skills. Digit. Journal. **10**(5), 866–887 (2022)
41. Johnson, P.R.: A case of claims and facts: automated fact-checking the future of journalism's authority. Digit. Journal. **1–24** (2023)
42. Hassan, N., Arslan, F., Li, C., Tremayne, M.: Toward automated fact-checking: detecting check-worthy factual claims by ClaimBuster. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1803–1812 (2017)
43. Atanasova, P., et al.: Automatic fact-checking using context and discourse information. J. Data Inf. Qual. (JDIQ) **11**(3), 1–27 (2019)
44. Guo, Z., Schlichtkrull, M., Vlachos, A.: A survey on automated fact-checking. Trans. Assoc. Comput. Linguist. **10**, 178–206 (2022)
45. Lecheler, S., Kruikemeier, S.: Re-evaluating journalistic routines in a digital age. New Media Soc. **18**(1), 156–171 (2016)
46. Lim, C.: Checking how fact-checkers check. Res. Polit. **5**(3), 2053168018786848 (2018)
47. Steensen, S., Kalsnes, B., Westlund, O.: The limits of live fact-checking: epistemological consequences of introducing a breaking news logic to political fact-checking. New Media Soc., 14614448231151436 (2023)
48. Nakov, P., et l.: Automated fact-checking for assisting human fact-checkers. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, Montreal, Canada, pp. 4826–4832. IJCAI (2021)
49. Müller, N., Wiik, J.: From gatekeeper to gate-opener: open-source spaces in investigative journalism. Journal. Pract. **17**, 189–208 (2023)
50. Powers, M.: In forms that are familiar and yet-to-be invented. Am. Journal. Discourse Technol. Specific Work. J. Commun. Inq. **36**(1), 24–43 (2012)

51. Olsen, G.R.: Enthusiasm and alienation: how implementing automated journalism affects the work meaningfulness of three newsroom groups. Journal. Pract., 1–17 (2023)
52. Lopez, M.G., Porlezza, C., Cooper, G., Makri, S., MacFarlane, A., Missaoui, S.: A question of design: strategies for embedding AI-driven tools into journalistic work routines. Digit. Journal. **11**(3), 484–503 (2023)
53. Dierickx, L., Lindén, C.G.: Journalism and fact-checking technologies: understanding user needs. Communication+1 **10**(1) (2023)
54. Samuelsen, R.J., Kalsnes, B., Steensen, S.: The relevance of technology to information verification: insights from norwegian journalism during a national election. Journal. Pract. **1–20** (2023)
55. Edwardsson, M.P., Al-Saqaf, W., Nygren, G.: Verification of digital sources in Swedish newsrooms-a technical issue or a question of newsroom culture? Journal. Pract. **17**(8), 1678–1695 (2023)
56. Weikmann, T., Lecheler, S.: Cutting through the hype: understanding the implications of deepfakes for the fact-checking actor-network. Digit. Journal. **1–18** (2023)
57. Reese, S.D.: Exploring the institutional space of journalism. Problemi dell Informazione **48**(1) (2023)
58. Pastor-Galindo, J., Nespoli, P., Mármol, F., Pérez, G.: The not yet exploited goldmine of OSINT: opportunities, open challenges and future trends. IEEE Access. **8**, 10282–10304 (2020)
59. Westlund, O., Larsen, R., Graves, L., Kavtaradze, L., Steensen, S.: Technologies and fact-checking: a sociotechnical mapping. In: Disinformation Studies: Perspectives from An Emerging Field, pp. 193–236. Labcom Communication & Arts, Covilhã, Portugal (2022)
60. Lindén, C.G.: What makes a reporter human? A research agenda for augmented journalism. Questions de communication **37**, 337–351 (2020)
61. Shkliarevsky, G.: The Emperor with No Clothes: Chomsky Against ChatGPT (2023). Available at SSRN 4439662
62. Larssen, U.: "But verifying facts is what we do!": fact-checking and journalistic professional autonomy. In: Democracy and Fake News: Information Manipulation and Post-Truth Politics, pp. 199–213. Routledge, London (2020)
63. Komatsu, T., et l.: AI should embody our values: investigating journalistic values to inform AI technology design. In: Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society, pp. 1–13, Association for Computing Machinery, New York (2020)
64. Schapals, A.K., Porlezza, C.: Assistance or resistance? Evaluating the intersection of automated journalism and journalistic role conceptions. Media Commun. **8**(3), 16–26 (2020)
65. Graves, L., Cherubini, F.: The rise of fact-checking sites in Europe. Digital News Project Report (2016)
66. Ward, S.J.A.: Global journalism ethics: widening the conceptual base. Glob. Media J. **1**, 137 (2008)
67. 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. Journal. **10**(10), 1775–1793 (2022)
68. Leiser, M.: Bias, journalistic endeavours, and the risks of artificial intelligence. In: Editor, F., Editor, S. (eds.) Artificial Intelligence and the Media, pp. 8–32. Edward Elgar Publishing, Cheltenham (2022)

69. Ferrario, A., Loi, M.: How explainability contributes to trust in AI. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 1457–1466. Association for Computing Machinery, New York (2022)
70. Jacovi, A., Marasović, A., Miller, T., Goldberg, Y.: Formalizing trust in artificial intelligence: prerequisites, causes and goals of human trust in AI. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 624–635. Association for Computing Machinery, New York (2021)
71. Lim, G., Perrault, S.T.: Explanation Preferences in XAI Fact-Checkers. European Society for Socially Embedded Technologies (EUSSET) (2022)
72. Micallef, N., Armacost, V., Memon, N., and Patil, S.: True or false: studying the work practices of professional fact-checkers. In: Proceedings of the ACM on Human-Computer Interaction, vol. 6, pp. 1–44. Association for Computing Machinery, New York (2022)
73. Nguyen, A.T., Kharosekar, A., Krishnan, S., Tate, E., Wallace, B.C., Lease, M.: Believe it or not: designing a human-AI partnership for mixed-initiative fact-checking. In: Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology, pp. 189–199. Association for Computing Machinery, New York (2018)
74. Demartini, G., Mizzaro, S., Spina, D.: Human-in-the-loop artificial intelligence for fighting online misinformation: challenges and opportunities. IEEE Data Eng. Bull. **43**(3), 65–74 (2020)
75. Hamed, A. A., Zachara-Szymanska, M., Wu, X.: Safeguarding authenticity for mitigating the harms of generative AI: Issues, research agenda, and policies for detection, fact-checking, and ethical AI. iScience **27**(2), 108782 (2024)
76. Van Witsen, A., Takahashi, B.: How science journalists verify numbers and statistics in news stories: towards a theory. Journal. Pract. **1–20** (2021)
77. Stray, J.: Making artificial intelligence work for investigative journalism. In: Thurman, N., Lewis, S.C., Kunert, J. (eds.) Algorithms, Automation, and News, pp. 97–118. Routledge, London (2021)
78. Montoro-Montarroso, A., et al.: Fighting disinformation with artificial intelligence: fundamentals, advances and challenges. Profesional de la información **32**(3) (2023)
79. Currie, G.M.: Academic integrity and artificial intelligence: is ChatGPT hype, hero or heresy? In: Seminars in Nuclear Medicine, pp. 1–13. Springer, Heidelberg (2023)
80. Ji, Z., et al.: Survey of hallucination in natural language generation. ACM Comput. Surv. **55**(12), 1–38 (2023)
81. Li, Z.: The dark side of chatGPT: legal and ethical challenges from stochastic parrots and hallucination. arXiv preprint arXiv:2304.14347 (2023)
82. Ray, P.P.: ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet Things Cyber-Phys. Syst. **3**(1), 121–154 (2023)
83. Yu, W.: A survey of knowledge-enhanced text generation. ACM Comput. Surv. **54**(11s), 1–38 (2022)
84. Kreps, S., McCain, R.M., Brundage, M.: All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. J. Exp. Political Sci. **9**(1), 104–117 (2022)
85. Brugger, F., Gehrke, C.: Skilling and deskilling: technological change in classical economic theory and its empirical evidence. Theory Soc. **47**, 663–689 (2018)

86. Polyportis, A., Pahos, N.: Navigating the perils of artificial intelligence: a focused review on ChatGPT and responsible research and innovation. Humanit. Soc. Sci. Commun. **11**(1), 1–10 (2024)
87. LaGrandeur, K.: The consequences of AI hype. AI Ethics, 1–4 (2023)
88. van Dalen, A.: Algorithmic Gatekeeping for Professional Communicators: Power, Trust, and Legitimacy. Taylor & Francis, London (2023)
89. Siau, K., Wang, W.: Building trust in artificial intelligence, machine learning, and robotics. Cutter Bus. Technol. J. **31**(2), 47–53 (2018)
90. Bartneck, C., Lütge, C., Wagner, A., Welsh, S.: Trust and fairness in AI systems. In: An Introduction to Ethics in Robotics and AI. SE, pp. 27–38. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-51110-4_4
91. Opdahl, A.L., et al.: Trustworthy journalism through AI. Data Knowl. Eng. **146**, 102182 (2023)
92. Deuze, M., Beckett, C.: Imagination, algorithms and news: developing AI literacy for journalism. Digit. Journal. **10**(10), 1913–1918 (2022)
93. Lopezosa, C., Codina, L., Pont-Sorribes, C., Vállez, M.: Use of generative artificial intelligence in the training of journalists: challenges, uses and training proposal. Profesional de la información/Inf. Prof. **32**(4) (2023)
94. Becker, K., et al.: Policies in parallel? A comparative study of journalistic AI policies in 52 Global News Organisations. *Oxford University Research Archive*, pp. 1–37 (2023)
95. Weisz, J.D., Muller, M., He, J., Houde, S.: Toward general design principles for generative AI applications. arXiv preprint arXiv:2301.05578 (2023)
96. Tonmoy, S.M., et al.: A comprehensive survey of hallucination mitigation techniques in large language models. arXiv preprint arXiv:2401.01313 (2024)
97. Feldman, P., Foulds, J.R., Pan, S.: Trapping LLM hallucinations using tagged context prompts. arXiv preprint arXiv:2306.06085 (2023)
98. Bsharat, S.M., Myrzakhan, A., Shen, Z.: Principled instructions are all you need for questioning LLaMA-1/2, GPT-3.5/4. arXiv preprint arXiv:2312.16171 (2023)
99. White, J., et al.: A prompt pattern catalog to enhance prompt engineering with chatgpt. arXiv preprint arXiv:2302.11382 (2023)
100. Rai, A.: Explainable AI: from black box to glass box. J. Acad. Mark. Sci. **48**, 137–141 (2020)
101. Weber-Wulff, D., et al.: Testing of detection tools for AI-generated text. Int. J. Educ. Integr. **19**(1), 26 (2023)