Tripodos, number 55 | 2024| 99-115
ISSN: 1138-3305

Received: 29/11/23
Accepted: 20/03/24
OnlineFirst: 24/06/24
Final publication: 15/07/24

Artificial Intelligence Tools and Bias in Journalism-related Content Generation: Comparison Between Chat GPT-3.5, GPT-4 and Bing

Mar Castillo-Campos
Loyola University Andalusia (Spain)
David Varona-Aramburu
Complutense University of Madrid (Spain)
David Becerra-Alonso
Loyola University Andalusia (Spain)

This study explores the biases present in artificial intelligence (AI) tools, focusing on GPT-3.5, GPT-4, and Bing. The performance of the tools has been compared with a group of experts in linguistics, and journalists specialized in breaking news and international affairs. It reveals that GPT-3.5, widely accessible and free, exhibits a higher tendency rate in its word generation, suggesting an intrinsic bias within the tool itself rather than in the input data. Comparatively, GPT-4 and Bing demonstrate differing patterns in term generation and subjectivity, with GPT-4 aligning more closely with expert opinions and producing fewer opinative words. The research highlights the extensive use of generative AI in media and among the general populace, emphasizing the need for careful reliance on AI-generated content. The findings stress the risks of misinformation and biased reporting inherent in unexamined AI outputs. The challenge for journalists and information professionals is to ensure accuracy and ethical judgment in content creation to maintain the quality and diversity of content in journalistic practices.

Keywords: *media bias, NLP, natural language, chat GPT, computational communication.*

Artificial intelligence (AI) is the branch of computing that allows machines to perform tasks that would normally require human intelligence (Abbott, 2010) and can learn from experience and progressively improve their performance without requiring explicit programming to do so (Dhiman,

Doi.org/10.51698/tripodos.2024.55.06

2023). In recent years, their improvement and use have been exponential in numerous fields, from medicine to journalism. Specifically for the latter area, AI based on Natural Language Processing (NLP) techniques is of particular concern. This technology can understand the language used by humans, process it and give a response in that same language code, whether through voice, text or images. Thus, it does not require structured inputs or information, something that had limited its use to users with that technical capacity. As these tools are able to understand unstructured information, they solve with enormous efficiency and give an answer, which opens the possibility of use to a much wider audience (Dwivedi *et al*., 2023). AI tools with natural language processing of generative texts, such as Chat GPT or Bing, reached the mass public at the end of 2022 and the beginning of 2023. ChatGPT, from the company OpenAI, is, in fact, the fastest-adopted technology in history. Estimates are 123 million monthly active users in less than three months after its launch. This surpasses TikTok, with 100 million monthly active users nine months after its release, or Instagram, which took two and a half years to reach the same figure (Wodecki, 2023). Other companies followed: Bard, by Google, which is currently not available in all countries, and Bing, a variant of OpenAI that Microsoft incorporated into its search engine also as a conversational chat (Gutiérrez-Caneda *et al*., 2023). OpenAI launched ChatGPT-4 on subscription in early 2023 and this is the latest update available at the time of this study. GPT-3.5 is still in use and with a much wider public reach, as it is free to use, as is the case with Bing. All of them coexist with other generative artificial intelligence tools, capable of creating images, videos, audio and other projects in a multitude of formats. Updates come fast, but the media —among many other industries— already use them to generate content, including news content (Türksoy, 2022), as well as for the elaboration of simple news items, the adaptation to different dimensions of the article (Gutiérrez-Caneda *et al*., 2023), the writing of headlines (Dale, 2021) or summaries (Goyal *et al*., 2022; Grail *et al*., 2021; Gupta *et al*., 2022; Liu and Healey, 2023), the detection of disinformation (Rai *et al*., 2022; Schütz *et al*., 2021), sentiment analysis (Leippold, 2023; Rathje *et al*., 2023), etc. The technological advancement of these tools and their widespread use by information professionals portends several perks that the media industry is already taking advantage of:

1. Increased text production, that is, also a wider range of topics that can be covered and at a lower cost (Dhiman, 2023; Noain-Sánchez, 2022; Türksoy, 2022).
2. Automation of non-specific content, allowing redirection of editors' efforts to more complex or specific coverage (Tejedor and Vila, 2021; Dalen, 2012).
3. More speed in the creation of texts for news events that require rapid publication, such as live events, emergencies, etc., and in different languages (Dhiman, 2023; Hassan and Albayari, 2022).

Other authors mention advantages such as accuracy due to the ability to quickly process a large amount of data (Dhiman, 2023; Noain-Sánchez, 2022), but this is highly dependent on the quality filter of that data, which may well contain

inaccuracies, biases or errors (Hassan and Albayari, 2022). Nevertheless, studies still warn of the need for human review of the machine's work (Bailer *et al.*, 2022; Dale, 2021). On the other hand, the use of generative AI in newsrooms also has some drawbacks:

1. This technology makes it possible to generate a large volume of content but does not ensure its quality. For users, these tools function as a sort of black box, so that in most cases it is not possible to trace the origin or processing of the data displayed (Barrio and Gatica-Pérez, 2023; Dwivedi *et al.*, 2023; Zhai, 2023). Users do not know what information constitutes the machine's knowledge, whether it uses, for example, personal, biased, false or sensitive data.
2. The technology does not have the capacity to distinguish good from bad, or ethical positions (Dale, 2021; Noain-Sánchez, 2022).
3. A third disadvantage is also put forward as the main hypothesis of this research (H1): Artificial intelligence systems, contrary to being neutral or unbiased ('aseptic'), inherently exhibit tendencies to generate content with embedded perspectives.

   As a result, the text produced by these systems can manifest an opinionated or biased orientation. Without careful oversight and mitigation strategies, this inherent bias has the potential to perpetuate misinformation.

This study has two further research objectives: (Objective 1) to check whether AI tools tend to generate new words or, on the contrary, extract them from the text they are given (in that case, the source of the bias could be more diffuse, because it could come from the input or the execution of the tool) and (Objective 2) to check which tool is more aligned to human experts.

Human-made news are partial by nature. Machine-generated news can have biases drawn from the human-made data they learn from. Based on this idea, in this study, we want to test whether these tools do indeed summarize information in an unbiased manner or whether they are tendentious. Some authors (Dhiman, 2023; Donk *et al.*, 2012; Gutiérrez-Caneda *et al.*, 2023; Kohring and Matthes, 2002) argue that the media's coverage of these technological tools is biased in a positive way. The analysis of Brennen and Nielsen (2018) considers that AI is generally presented as a solution to practically all kinds of problems. It is possible that the image of this type of tool being projected is particularly positive in terms of professional uses and neglects to reinforce some of the shortcomings or risks it still has. There are a variety of reflections and ethical studies on the subject (e.g. Hurlburt, 2023; Niederman and Baker, 2023; Zohny *et al.*, 2023). In any case, the regulation of AIs is currently evolving much more slowly than the technologies themselves, and there is evidence of perverse uses of these tools by many users (Verma, 2023). The industry itself, except for the AI development companies mentioned above, issued a manifesto in March 2023 calling for a pause in the training of these technologies, arguing that the legal, ethical and social frameworks are unclear, and excesses may be committed (Vincent, 2023). Beyond self-regulation, measures are taken on an ad hoc basis and limited to

specific environments or geographic areas. For example, the Italian government announced that it would ban the use of Chat GPT in its territory for infringing data protection, but this ban was not formalized (Rodríguez de Luis, 2023). Similarly, many universities in Europe are regulating their use internally (Carabantes *et al*., 2023). Currently, and until legislation creates stable frameworks for the regulation of their use, the use of these tools must be conscious of the risks involved and the negative effects they may generate. As some authors note, there is a great challenge in the journalism industry to train professionals in AI and its concrete tools (Gonçalves and Melo, 2022; Stray, 2019).

NLP TOOLS AND INFORMATION BIAS

This work builds on the consensus of academia and news professionals on the essential difficulty of achieving journalistic objectivity. As Whittaker (2019) notes objectivity has been considered an essential principle of journalism and has become a myth thoroughly debunked and discredited among media theorists (Knight and Cook, 2013). Nevertheless, the pursuit of objectivity "remains firmly embedded in the professional practice of journalism, and the more the profession is criticized, the more objectivity is defended as a necessary part of the contribution news organizations make to society as a whole" (Knight and Cook, 2013). However, many authors speak of a new era of journalism, in which the boundaries between opinion and fact, information and entertainment are blurring, changing "the traditional rules of political communication," which are no longer valid (Llorca-Abad and López-García, 2020).

Bias research is approached from a variety of perspectives: sentiment analysis (Taboada, 2016), topic modelling (Kherwa and Bansal, 2019) or framing (Scheufele, 1999), among others. All of them start from the premise that all information carries an implicit bias, personal and unavoidable, committed by human beings, from the decision to cover a topic to the way it is written, or the hierarchy given to the news in the media in which it is published. While AI tools themselves do not register feelings or preferences, the data they are trained with comes from humans, who may have consciously or unconsciously biased that information. Even if the tool is unbiased, the results provided by the tool would not necessarily be so, as the training data could be biased from the start. The more diverse the view of the documents that make up the training corpus, the more ideologically wide-ranging the response of the machine will also be. However, users of these tools can't know which texts or postures are used in training.

Some publications report readers' opinions that news stories written only by generative AI seem more accurate and objective (Clerwall, 2017; Dalen, 2012). In fact, one of the risks currently posed is that opacity in the sources and structure of information "contributes to creating an image of trustworthiness and honesty, [which] makes regulation of these tools especially necessary to avoid further disinformation" (Gutiérrez-Caneda *et al*., 2023). However, to date, we have not found studies that effectively demonstrate that the most widespread AI tools

generate biased or, moreover, opinative information, we have proceeded to make a comparison between three of the most widely used technologies: GPT- 3.5, GPT-4 and Bing. This research aims to demonstrate through the generation of content and the comparison between the different technologies that the words or texts generated may in themselves contain a bias or opinative trait that is important for journalists to consider. The first approaches (Gutiérrez-Caneda *et al*., 2023) reflect a certain concern among journalists, who are aware that they could start with biased or false information or put the user's privacy at risk by using their information as a learning tool.

In any case, the interest of this study lies in the growing use and scope of these tools. Moreover, it should be understood as an exploratory work that tries to frame the current situation. It is understood that these are evolving tools and that, therefore, it is possible that their results will also change in the future. As indicated in other studies in the area, the changes brought about by these programs are recent and changing, but with a very significant impact, so "it is necessary to monitor them periodically and as up to date as possible in order to understand the trends and to know the most commonly used software at the present time" (Gutiérrez-Caneda *et al*., 2023). Thus, we consider that the intrinsic bias in these tools is a relevant point to be considered by users, professionals and academics who use them today. This research arises from this conviction.

METHODOLOGY

This is experimental research with mixed methodologies. The headlines were subjected to an unsupervised automatic classification and then cross-checked with the answers provided by a panel of experts.

Several prompts have been chosen as a common question for the three technologies: GPT-3.5, GPT-4 and Bing. These prompts are instructions that guide the machine in the generation of answers or actions. They have been refined according to the conciseness of the answer until finding the one that best fits the type of result expected. The answers have been obtained through the conversational chat offered by the tools and have been collected in a table of contents. On the other hand, an identical questionnaire was passed to a group of 9 journalists and academic experts with a limited selection of these headlines (n=20 out of the initial sample of n=199). With this methodological crossover, it is hoped that differences between automatic and manual generation, if any, can be verified.

For the experiment, news headlines from two Spanish media outlets, *El Mundo* and *20 minutos*, were used. The headlines correspond to all the news items (N=184) that these media outlets have labelled as relating to the Ukraine-Russia conflict during the first month of media coverage of the war in 2022. The sample also includes 15 other random news headlines dealing with topics other than the conflict, published by the same media outlets during the same period (Figure 1).

Figure 1. Description of tasks in the workflow

Phase	Task
Data extraction	Headlines extracted from the digital newspaper repository. All news items labelled by the media themselves as referring to the Ukraine-Russia conflict during the first month of the war (February 24 - March 24 2022). Other random non war-related headlines published by the same media in the same time period are added.
Natural language processing	Realised with GPT-3.5, GPT-4 and Bing. Generation of summary words from headlines (H1). AI tools were asked about the argument for their decision. They were asked to determine whether these words were positive, negative or neutral (H1)
Group of experts	They were asked to perform the same tasks as AI tools, but on a smaller sample (O2)
Manual separation of summary words (O1)	Whether they were extracted from the text (appear in the headline) or generated (do not appear) (O1). On the words generated by the tools: whether they are positive, negative or neutral is contrasted. Comparison with the REDES dictionary and the tools' own classification (H1).
Data analysis	The performance of the three tools is contrasted: the generation of words with tendency and the comparison with the group of experts

Source: Own elaboration.

Then, a search was made to obtain a summary word for each given headline. Once all the terms had been collected, a distinction was made between those extracted from the text, i.e., which were already present in the headline and the AI selected them as a keyword; and those generated by the tool, which did not appear in the given text and were therefore proposed by the AI. From the latter group, those with a connotative meaning have been manually selected. To ascertain their connotative significance, the identified terms underwent scrutiny against the lexical resource *REDES: Diccionario combinatorio del español contemporáneo* (Bosque, 2004), renowned for furnishing contextual usage nuances for individual words. Additionally, validation was performed utilizing GPT-3.5, GPT-4, and Bing tools.

The lexical resource under scrutiny furnishes an array of lexemes associated or co-occurring with the queried term. It has been discerned that all examined terms: a) are employed subjectively, or b) demonstrate a propensity to be conjoined with lexemes bearing positive or negative connotations. Certain lexemes exhibit overtly pejorative undertones ('aberrant', for instance), while others convey positivity ('goodness'), and yet others tend to be associated with either positive or negative lexemes, contingent upon the term ('failure', 'defiance'). A subset, regarded as neutral, is paired with relational or determinative lexemes, or with equivalent probability, positive and negative lexemes ('impact').

Secondly, GPT-3.5, GPT-4, and Bing were utilized to assess the sentiment (positive, negative, or neutral) of the terms. Terms that were deemed neutral by all three technologies were excluded from further analysis. This sentiment analysis was exclusively applied to terms generated by the AI, rather than those extracted from the headline. This approach is predicated on the understanding that biases may emanate not only from the initial input (e.g., the headline) but also from the inherent biases within the AI tool itself. In contrast, terms

extracted directly from sources may have a more ambiguous origin. By focusing on AI-generated terms, this methodology aims to isolate and identify potential biases embedded in the AI's training data.

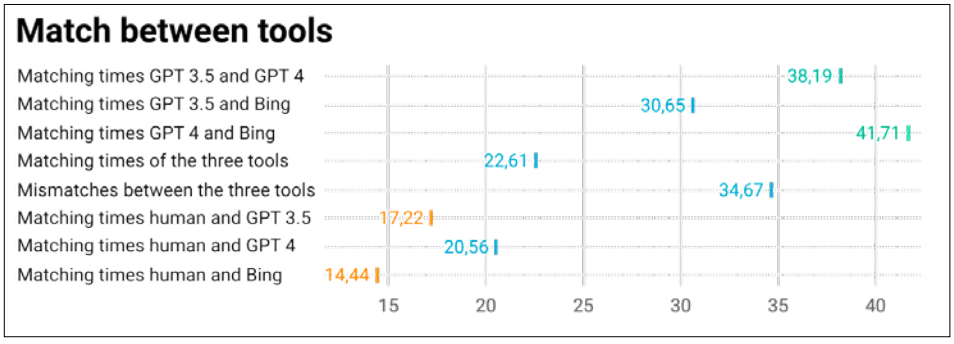
RESULTS

Given a headline, and requested from it a summary word, Bing is the technology that in a higher percentage chooses a word from the text itself, in 35.68% of cases (10 percentage points more than the other two tools). However, on most occasions, the tools generate a word that is absent from the given text: GPT- 3.5 does so 74.87% of the time, GPT-4 75.38% and Bing 64.32%. In contrast, experts tend to choose a word from the given text itself in 31.28% of cases and generate their own word summary in 68.72% of cases, which does not differ much from the results obtained with AI.

On 22.61% of occasions, the three technologies match the resulting word. GPT-4 and Bing match in 41.72% of the cases, the highest percentage in pairwise matches. In 34.67% of cases, none of the three matched. The group of experts, on the other hand, showed agreement with one of these AI tools 17.4% of the time. The experts agree 20.56% of the time with GPT-4, 17.22% with GPT-3.5, and 14.44% with Bing.

It is more likely, for example, that all three tools will agree on the result than that any of them will agree with the experts. The probability of agreement between GPT-4 and Bing is substantially higher than between GPT-3.5 and Bing. Experts are more likely to agree with GPT-4 than with the other AI technologies (Figure 2).

Figure 2. Percentage of match of summary words between tools and with the expert panel



Source: Own elaboration.

On the other hand, looking at the words generated by the tool itself, it is detected that some of them carry an implicit opinion. Terms such as 'hypocrisy', 'corruption', 'aberrant' or 'love', to cite a few examples, have been detected,

which are not merely descriptive or informative words. There are 26 words that could go beyond mere descriptiveness, and these appear up to 42 times.

Our analysis shows that of all the words generated by AI, 7.04% would have a subjective charge. GPT-3.5 would have generated 8.54% of words with connotations: 'solidarity', 'tragedy', 'rescue', 'coherence', 'boycott', 'effectiveness', 'sacrifice', 'inspiring', 'revealing' and 'controversy'. GPT-4 would have 5.53%: 'plea', 'solidarity', 'contradiction', 'setback', 'tragedy', 'failure', 'challenge' and 'rescue'. Bing, 7.04%: 'hypocrisy', 'corruption', 'aberrant', 'determination', 'tragedy', 'love', 'ambition', 'solidarity', 'overcoming', 'impossible', 'compassion' and 'supplication'.

This determination has been contrasted with the context provided by the dictionary *REDES: Diccionario combinatorio del español contemporáneo* for each term, and the tools themselves (GPT-3.5, GPT-4 and Bing) have been asked for the positive, negative, or neutral denotation of each term (discarding the neutral ones) (Figure 3). With consensus among the sources consulted, the following data is obtained from the AI.

Figure 3. Example of terms generated by the three technologies. The usual linkage with terms according to the REDES dictionary and the classification as a word with positive, negative or neutral connotation according to GPT-3.5, GPT-4 and Bing are given

Term	Linked to (REDES)	GPT-3.5 classification	GPT-4 classification	Bing classification
Aberration	atrocious, appalling, flagrant, horrendous, unforgivable, inadmissible, dangerous, sinister...	negative	negative	negative
Ambition	blind, compulsive, unbridled, insane, unconscionable, excessive, unhealthy. Honest, legitimate, natural. Eagerness of desire	neutral	neutral	neutral
Love	affection, friendship, love affection, romance, tenderness	positive	positive	positive
Coherence	absolute, admirable, overwhelming. Combined with verbs (...) denoting choice or resolution, (...) or analysis, reflection, (...) to undertake a task or face a difficult task	positive	positive	positive
Corruption	to denounce, to extirpate, serious, scandal (of), vice... / with things that can cause physical or moral suffering and with nouns that designate feelings of irritation, resentment, animosity...	negative	negative	negative

Term	Linked to (REDES)	GPT-3.5 classification	GPT-4 classification	Bing classification
Defiance	authentic, true, certain, difficult, frontal, historical, important. Threat, challenge, risk (of)	positive	positive	positive
Determination	right, categorical, clear, decisive, free. Choice, bravery, courage	positive	positive	positive
Failure	absolute, bitter, demolishing, honourable. Fall, defeat, disaster, disappointment	negative	negative	negative
Persistence	is constructed with nouns that designate (...) verbal statements that are assertive, declarative or against someone. Particularly outstanding are its combinations with nouns that denote failure, mistake (...), to break a law, to disrespect...	positive	positive	positive
Resistance	heroically, tenaciously, bravely. Maintenance, opposition, rejection	positive	positive	positive
Solidarity	enormous, spontaneous, generous, necessary, responsible. Supportive, understanding, united...	positive	positive	positive
Overcoming	achievement, success	positive	positive	positive
Tragedy	bitter, distressing, Dantesque, desolate, harsh. Drama, fatality	negative	negative	negative
Impossible	absolutely, in every way	negative	negative	neutral
Backlash	alarming, clear. Advance, regression	negative	negative	neutral
Contradiction	absurdity, contradiction, contrast, mistake, opposition	negative	negative	neutral
Compassion	affliction, pain, suffering, feeling	positive	positive	neutral

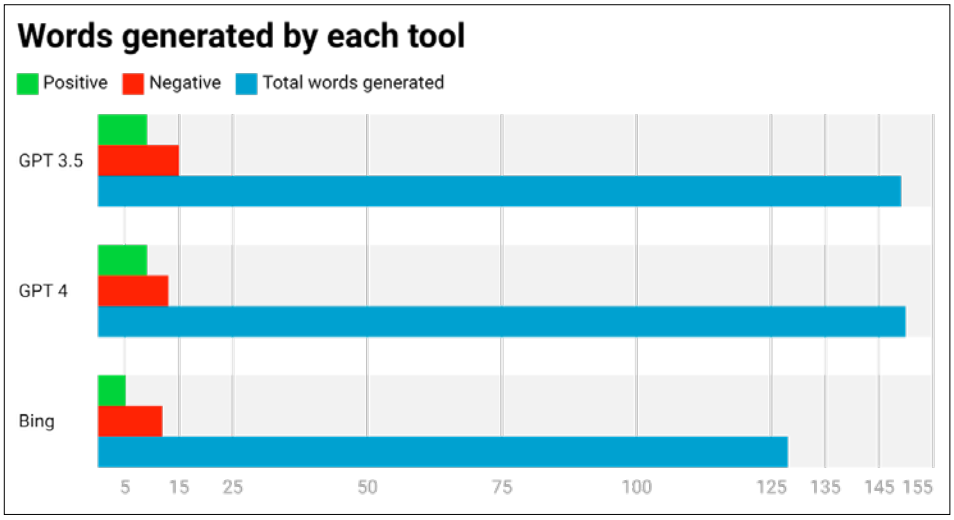
Source: Own elaboration.

Of these 26 words, chat GPT-3.5 classifies 9 terms as negative ('hypocrisy', 'corruption', 'aberrant', 'tragedy', 'failure', 'impossible', 'backwardness', 'contradiction' and 'boycott') and 15 positives ('determination', 'persistence', 'endurance', 'love', 'solidarity', 'rescue', 'coherence', 'challenge', 'overcoming', 'compassion', effectiveness', 'supplication', 'sacrifice', 'inspiring' and 'revealing').

GPT-4 considers these same 9 words to have negative connotations and classifies 13 as positive ('determination', 'persistence', 'endurance', 'love', 'solidarity', 'consistency', 'challenge', 'overcoming', 'compassion', 'effectiveness', 'sacrifice', 'inspiring' and 'revealing'). For Bing, only 5 words are negative ('hypocrisy', 'corruption', 'aberrant', 'tragedy' and 'failure') and 12 are positive

('determination', 'persistence', 'resilience', 'love', 'solidarity', 'rescue', 'coherence', 'challenge', 'overcoming', 'sacrifice', 'inspiring' and 'revealing') (Figure 4).

Figure 4. Positive or negative words among those generated by each tool



Source: Own elaboration.

Some examples are given below. Given the prompt "Summarize this headline in one word," and the input "Putin, just 10 days ago, to German Chancellor: 'Do we want war? Of course, not'" (from the media outlet *20 minutos*), the summary word generated by Bing is 'Hypocrisy' and argues: "Because Putin says he doesn't want war, but at the same time he supports authoritarian regimes and violates human rights". The information that the technology claims to use as an argument to produce the word summary does not appear in the headline and is, in any case, a qualifying adjective assigned to the subject. The word provided by GPT-3.5 is imprecise: 'No'. For GPT-4, 'contradiction'. The latter, although it could imply an opinative charge, could also be due to the nuance of the word 'just'.

To the same question, but with the headline "Putin bought a luxury villa in Biarritz in 1996, when he was still working as a public employee" (*20 minutos*), Bing summarizes with 'Corruption' and offers the argument: "Because Putin acquired a high-value property when he had a modest salary as a civil servant". Again, the information used by the machine is not explicit in the text provided and seems to exceed the bounds of objectivity, since there may be other reasons why the subject could have acquired the property, for example. For GPT-3.5, the keyword is 'Putin'; 'Villa' for GPT-4.

Asked the same question to the interrogative headline "Could Russia's aggression against Ukraine have been prevented?" (from *El Mundo*), Bing's answer is 'Impossible', which is an opinionated response to the rhetorical question, not a summary of the headline or a keyword in the headline. Chat GPT generates 'Prevention' in both versions.

Regarding news not related to the Ukraine-Russia conflict, the first example is "Thousands of child porn videos found in Florida foster father's home" (*20 minutos*). The word provided by Bing is 'Aberrant' and the explanation given: "This word means something that deviates from what is normal or morally acceptable and reflects the gravity and horror of what the foster father did. Other words that might serve are 'heinous', 'disgusting' or 'monstrous'". While this argument may reflect a generalized thought or common value, it is nonetheless opinative, in comparison also to the keyword generated by GPT-3.5, 'paedophilia', which reflects the motive that could have caused the news story.

For the headline "Yolanda Díaz will launch her project in April and gives herself six months to build it without losing sight of the Galician political tides", Bing generates 'Ambition', under the argument that "the headline implies that Yolanda Díaz has her own political project that she wants to develop and consolidate in a given period of time, without neglecting her influence in Galicia. Ambition is the desire to achieve something important or difficult." Depending on the political positioning of the reader, this headline could indeed seem to reflect ambition or, otherwise, reprehensible behavior. GPT-3.5 does not assess the text and concludes with the word 'Marés', which claims to refer to the Galician political tides. GPT-4 concludes 'Project'.

Through expert judgment, it has been possible to compare the responses to the same questions posed to AI and nine journalists or experts in conflict communication. The experts often provide a summary word that is not present in the headline, only resorting to words already in the headline 31.28% of the time. The highest rate of agreement among the experts is with GPT-4 (Figure 5).

Figure 5. Summary words generated by the experts or selected from the headlines and their agreement with AI

Expert	Coincidence with GPT-3.5	with GPT-4	with Bing	words extracted	words generated
AT	3	2	3	2	18
AS	1	1	2	2	18
AL	4	5	1	4	16
IA	5	6	3	8	12
OG	6	6	6	13	7
PH	4	5	4	19	1
MB	1	2	3	1	19
MHR	1	2	1	0	19
SH	6	8	3	7	13
TOTAL	31	37	26	56	123
Percentage	17,22%	20,56%	14,44%	31,28%	68,72%

Source: Own elaboration.

As part of the research, we wanted to analyze whether the AI tools generated or, on the contrary, extracted a word from the given text when given a headline or sentence to analyze and asked for a summary word (Objective 1). On most occasions, they did generate it (in 71.52% of cases, on average), slightly more often than experts, who did it in 68.72% of cases.

When using these automated tools, about 7% of the time, the words they generate show some bias or tendency, and in the case of GPT-3.5, this is 8.54%. The fact that these percentages correspond to the words generated by the tool and not to those extracted directly from the given text suggests that, if a bias exists, it is produced in the tool and not in the inputs of the experiment. These data have been taken with caution, as it is understood that many words are not positive or negative per se and require more detailed context. However, it is the tools themselves that indicate that for most of the words with detected connotations, there is a negative and/or positive loading of the terms. Therefore, it is accepted that AI systems, contrary to being neutral or unbiased, inherently exhibit tendencies to generate content with embedded perspectives (H1).

The results of GPT-4 and Bing match 41.72% of the time when generating identical words. So far, Bing is known to integrate some of the GPT-4 technology, although there are some differences in term generation. GPT-4 tends to generate words at ten percentage points higher than Bing, which gets a higher rate in extracting a word from the given text. Thus, Bing generates 7.04% of words with a certain subjectivity or tendency (compared to 5.53% for GPT-4), although it does not detect in the same way whether they are positive or negative words (it detects 17 terms with connotation, compared to 22 for GPT-4 out of a total of 26 words). In turn, experts agree more often with GPT-4 than with Bing (Objective 2). In any case, GPT-3.5 is the tool that records the fewest coincidences with its counterparts and the one with the highest rate of generated subjectivity (8.54% of the words). It is noteworthy that Chat GPT-3.5 is offered in the free version at the time of this study and is the most widely used by users.

DISCUSSION

The employment of generative artificial intelligence in newsrooms and the media sector is currently a reality. Its utilization extends to the general public as well, marking a technology that is witnessing an unparalleled surge in user adoption and is increasingly being applied across various domains. While its effectiveness is established, it is equally important to acknowledge the risks associated with blindly trusting AI-generated content, as highlighted by previous scholars (Bailer *et al.*, 2022; Dale, 2021).

Given the opacity of the training data, it is challenging to foresee which subjects might exhibit particular biases, be outdated, or lack comprehensiveness. Although generative AI serves as a valuable tool, journalists have a responsibility to ensure accuracy in their reporting, and uncritically accepting AI-generated content could lead to the dissemination of misinformation. Journalists might

inadvertently propagate biased perspectives or stereotypes, which could impact the fairness and balance of their reporting. This is especially important when covering events with complex political or cultural contexts because AI models may lack in-depth knowledge of the facts and current events. This could result in content that is superficial or misses nuanced perspectives critical in journalism. Factual analysis should always be critical, subject to the ethical judgement of the journalist and published with an awareness of where the information comes from, how reliable it is and what impact it may have on readers. In addition, there is a risk of generating content that closely resembles existing material, leading to plagiarism concerns.

As introduced by other scholars, the boundaries between information and entertainment, as well as between opinion and information, are becoming increasingly blurred (Llorca-Abad and López-García, 2020). AI tools may be contributing to this shift, as AI-generated text has been identified as commonly biased or opinionated in this study. Regarding our research objectives, we explored whether the most used tools align with those that experts most frequently endorse (Objective 2). Tools that are offered at no cost to the user and are more widely used by the general population also tend to show a higher degree of bias. While it is difficult to obtain precise data on the percentage of users opting for GPT-4 over GPT-3.5, it is estimated at 1% (Nerdynav, 2023). It is challenging to determine whether the bias stems from the inputs or the tool itself (Objective 1), but the percentage of generated words could be indicative of the AI's tendency to propose biased content. Again, this issue is more pronounced in free and more widely used tools.

This research is only intended to draw the attention of information professionals to encourage human control over artificially generated texts and to promote the study of the quality and plurality of the content generated with these tools. Given that their use is increasingly widespread among users and newsrooms, we wanted to explore how these technologies are related to current issues and for which AI is likely to be used in newsrooms, such as the Ukraine-Russia conflict. It is crucial to understand AI's functionality and associated risks, as also suggested by previously mentioned scholars (Gonçalves and Melo, 2022; Stray, 2019).

It should be borne in mind that the main limitation is that technology is advancing extremely fast. This study aims to regularly monitor these updates to understand how these tools work, comprehend their mechanisms, and learn about the risks and benefits associated with their use. OpenAI recently released a bias-checker tool, which is still in beta but is likely to contribute to bias detection in the near future. Other investigations should make use of such tools. Future research will include longer texts, which will also demand a more in-depth analysis of the text generated by AI. The generation and analysis of single words makes it difficult, on the one hand, to evaluate without context; but it also requires a clear positioning. In many cases, this inclination meant, for experts and machines, a preference for opinion. In any case, the exploratory nature of this article suggests some future research and aspires to contribute to the journalistic field as far as it warns of the generation of biased or opinionated responses.

CONCLUSION

This research has critically assessed the capabilities of GPT-3.5, GPT-4, and Bing in generating content, with a particular emphasis on identifying any inherent opinionated components or biases. By employing a dataset comprising news reports on a specific military conflict and benchmarking the outcomes against the insights of a panel of experts, this study offers insights into the implications of using advanced AI tools in the dissemination and interpretation of complex events. The hypothesis "Artificial intelligence systems, contrary to being neutral or unbiased, inherently exhibit tendencies to generate content with embedded perspectives" is accepted. The findings underscore the necessity of cautious engagement with these technologies, highlighting their potential impacts on the accuracy and impartiality of information shared in the public domain.

GPT-3.5 shows a greater propensity for certain word generation, suggesting a potential intrinsic bias embedded within the tool, as opposed to biases in the input data. In contrast, GPT-4 and Bing exhibit distinct patterns in terms of word selection and subjectivity, with GPT-4 demonstrating a closer alignment with expert viewpoints and generating a reduced number of subjective terms. We present several examples where we observe how, from a news headline (input) from a media outlet, the summary word generated by AI becomes laden with connotation, typically negative. We distinguish between summary words that appear in the input itself and those summary words that do not appear in the input and are, therefore, generated (not extracted) by the AI itself.

The implications of these findings extend beyond the immediate sphere of journalism. They touch on the broader usage of AI-generated content across media and by the general populace, underscoring the urgent need for a critical and informed engagement with AI tools. The risk of misinformation and biased reporting, as highlighted in this study, serves as a reminder of the responsibility borne by journalists and information professionals. A false assumption of objectivity of AI tools can exacerbate the risk, leading journalists to neglect the oversight of generated content. It underscores the necessity of employing accuracy, ethical judgment, and a deep understanding of these tools' underlying mechanics in content creation. The difficulty of reaching a consensus on what bias is and how to manage it in journalism is understood. However, the issue is that bias is no longer merely a result of only human decision-making in coverage choices, writing, or prioritization. It has also become a, perhaps unconscious, replication by machines that we use without understanding their inherent biases.

Mar Castillo-Campos has a degree in Communication, and a master's degree in Research Methods. She is currently working at Loyola University Andalusia as a research assistant, integrating quantitative methodologies and data science in the field of journalism.

David Varona-Aramburu has a PhD in Journalism and currently works at the Department of Journalism and New Media at Complutense University of Madrid. David has worked for more than 20 years in the media, and does research in Communication and Media.

David Becerra-Alonso obtained his PhD in the School of Computing at the University of the West of Scotland, where he worked on dynamical chaotic systems. David currently holds a position as a lecturer at Loyola University Andalusia. His research interests include dynamical systems, emergent collective behavior, and machine learning techniques and heuristics.

References

Abbott, Andrew (2010). *Chaos of disciplines*. University of Chicago Press.

Bailer, Werner; Thallinger, Georg; Krawarik, Verena; Schell, Katharina, and Ertelthalner, Victoria (2022). AI for the media industry: Application potential and automation levels. In *International Conference on Multimedia Modeling* (pp. 109-118). Springer International Publishing. https://doi.org/10.1007/978-3-030-98358-1_9

Barrio, David Alonso del and Gatica-Pérez, Daniel (2023). Framing the news: From human perception to large language model inferences. *arXiv preprint arXiv:2304.14456*.

Bosque, Ignacio (2004). *Redes: Diccionario combinatorio del español contemporáneo. Las palabras en su contexto*. Editorial SM.

Brennen, J. Scott and Nielsen, Rasmus Klein (2018). *An industry-led debate: How UK media cover artificial intelligence*. Reuters Institute for the Study of Journalism. https://doi.org/10.60625/risj-v219-d676

Carabantes, David; González-Geraldo, José L., and Jover, Gonzalo (2023). ChatGPT could be the reviewer of your next scientific paper: Evidence on the limits of AI-assisted academic reviews. *Profesional de la Información*, *32*(5).

Clerwall, Christer (2017). Enter the robot journalist: Users' perceptions of automated content. In *The Future of Journalism: In an Age of Digital Media and Economic Uncertainty* (pp. 165-177). Routledge.

Dale, Robert (2021). GPT-3: What's it good for? *Natural Language Engineering, 27*(1), 113-118. https://doi.org/10.1017/S1351324920000601

Dalen, Arjen van (2012). The algorithms behind the headlines: How machine-written news redefines the core skills of human journalists. *Journalism Practice, 6*(5-6), 648-658. https://doi.org/10.1080/17512786.2012.667268

Dhiman, Bharat (2023). Does artificial intelligence help journalists: A boon or bane? https://dx.doi.org/10.2139/ssrn.4401194

Donk, André; Metag, Julia; Kohring, Matthias, and Marcinkowski, Frank (2012). Framing emerging technologies: Risk perceptions of nanotechnology in the German press. *Science Communication, 34*(1), 5-29. https://doi.org/10.1177/1075547011417892

Dwivedi, Yogesh K.; Kshetri, Nir; Hughes, Laurie; Slade, Emma L.; Jeyaraj, Anand; Kar, Arpan K.; Baabdullah, Abdullah M.; Koohang, Alex; Raghavan, Vishnupriya; Ahuja, Manju; Albanna, Hanaa; Albashrawi, Mousa A.; Al-Busaidi, Adil S.; Balakrishnan, Janarthanan; Barlette, Yves; Basu, Sriparna; Bose, Indranil; Brooks, Laurence; Buhalis, Dimitrios,... Wright, Ryan (2023).

"So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management, 71*, 102642.

Gonçalves, Adriana and Melo, Paulo Victor (2022). Inteligencia artificial y periodismo: Una aproximación al contexto portugués. *Fonseca, Journal of Communication*, (25), 23-34. https://doi.org/10.14201/fjc.29682

Goyal, Tanya; Li, Junyi Jessy, and Durrett, Greg (2022). News summarization and evaluation in the era of GPT-3. *arXiv preprint arXiv:2209.12356*.

Grail, Quentin; Pérez, Julien, and Gaussier, Eric (2021). Globalizing BERT-based transformer architectures for long document summarization. In *Proceedings of the 16th conference of the European chapter of the Association for Computational Linguistics: Main volume* (pp. 1792-1810). http://dx.doi.org/10.18653/v1/2021.eacl-main.154

Gupta, Anushka; Chugh, Dishka; Anjum, and Katarya, Rahul (2022). Automated news summarization using transformers. In *Sustainable advanced computing* (pp. 249-259). Springer. https://doi.org/10.48550/arXiv.2108.01064

Gutiérrez-Caneda, Beatriz; Vázquez-Herrero, Jorge, and López-García, Xosé (2023). AI application in journalism: ChatGPT and the uses and risks of an emergent technology. *Profesional de la Información, 32*(5).

Hassan, Abdulsadek and Albayari, Akram (2022). The usage of artificial intelligence in journalism. In *Future of organizations and work after the 4th industrial revolution: The role of artificial intelligence, big data, automation, and robotics* (pp. 175-197). Springer. http://dx.doi.org/10.1007/978-3-030-99000-8_10

Hurlburt, George (2023). What if ethics got in the way of generative AI? *IT Professional, 25*(2), 4-6. https://doi.ieeecomputersociety.org/10.1109/MITP.2023.3267140

Kherwa, Pooja and Bansal, Poonam (2019). Topic modeling: A comprehensive review. *EAI Endorsed Transactions on Scalable Information Systems, 7*(24). https://doi.org/10.4108/eai.13-7-2018.159623

Knight, Megan and Cook, Clare (2013). *Social media for journalists: Principles and practice*. Sage.

Kohring, Matthias and Matthes, Jörg (2002). The face(t)s of biotech in the nineties: How the German press framed modern biotechnology. *Public Understanding of Science, 11*(2), 143. https://doi.org/10.1088/0963-6625/11/2/304

Leippold, Markus (2023). Sentiment spin: Attacking financial sentiment with GPT-3. *Finance Research Letters*. https://doi.org/10.1016/j.frl.2023.103957

Liu, Sengjie and Healey, Christopher G. (2023). Abstractive summarization of large document collections using GPT. *arXiv preprint arXiv:2310.05690*.

Llorca-Abad, Germán and López-García, Guillermo (2020). Communication flows in the European elections: Amid populism and Euroscepticism. *Tripodos*, (49), 9-12.

Nerdynav (23 November 2023). 107 up-to-date CHATGPT statistics & user numbers [Nov 2023]. *Nerdynav*. https://nerdynav.com/chatgpt-statistics/

Niederman, Fred and Baker, Elizabeth White (2023). Ethics and AI issues: Old container with new wine? *Information Systems Frontiers, 25*(1), 9-28.

Noain-Sánchez, Amaya (2022). Addressing the impact of artificial intelligence on journalism: The perception of experts, journalists, and academics. *Communication & Society, 35*(3), 105-121. https://doi.org/10.15581/003.35.3.105-121

Rai, Nishant; Kumar, Deepika; Kaushik, Naman; Raj, Chandan, and Ali, Ahad (2022). Fake news classification using transformer based enhanced LSTM and BERT. *International Journal of Cognitive Computing in Engineering, 3,* 98-105.

Rathje, Steve; Mirea, Dan-Mircea; Sucholutsky, Ilia; Marjieh, Raja; Robertson, Claire, and Bavel, Jay J. van (2023). GPT is an effective tool for multilingual psychological text analysis. https://doi.org/10.31234/osf.io/sekf5

Rodríguez de Luis, Eva (2023). Chatgpt vuelve a Italia: Qué medidas ha implementado OpenAi para levantar el veto. *Genbeta*. https://www.genbeta.com/actualidad/chatgpt-vuelve-a-italia-que-medidas-ha-implementado-openai-para-levantar-veto

Scheufele, Dietram A. (1999). Framing as a theory of media effects. *Journal of Communication, 49*(1), 103-122. https://doi.org/10.1111/j.1460-2466.1999.tb02784.x

Schütz, Mina; Schindler, Alexander; Siegel, Melanie, and Nazemi, Kawa (2021). Automatic fake news detection with pretrained transformer models. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10-15, 2021, Proceedings, Part VII* (pp. 627-641). Springer International Publishing. https://doi.org/10.1007/978-3-030-68787-8_45

Stray, Jonathan (2019). Making artificial intelligence work for investigative journalism. In *Algorithms, Automation, and News* (pp. 97-118). https://doi.org/10.1080/21670811.2019.1630289

Taboada, Maite (2016). Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics, 2*, 325-347. https://doi.org/10.1146/annurev-linguistics-011415-040518

Tejedor, Santiago and Vila, Pere (2021). Exo journalism: A conceptual approach to a hybrid formula between journalism and artificial intelligence. *Journalism and Media, 2*(4), 830-840. https://doi.org/10.3390/journalmedia2040048

Türksoy, Nilüfer (2022). The future of public relations, advertising, and journalism: How artificial intelligence may transform the communication profession and why society should care*? Türkiye İletişim Araştırmaları Dergisi*, (40), 394-410. https://doi.org/10.17829/turcom.1050491

Verma, Pranshu (5 November 2023). AI fake nudes are booming. It's ruining real teens' lives. *The Washington Post*. https://www.washingtonpost.com/technology/2023/11/05/ai-deepfake-porn-teens-women-impact/

Vincent, James (29 March 2023). Elon Musk and top AI researchers call for pause on "giant AI experiments". *The Verge*. https://www.theverge.com/2023/3/29/23661374/elon-musk-ai-researchers-pause-research-open-letter

Whittaker, Jason Paul (2019). *Tech giants, artificial intelligence, and the future of journalism*. Taylor & Francis. http://library.oapen.org/handle/20.500.12657/25879

Wodecki, Ben (3 February 2023). UBS: ChatGPT may be the fastest growing app of all time. *AI Business*. https://aibusiness.com/nlp/ubs-chatgpt-is-the-fastest-growing-app-of-all-time

Zhai, Xiaoming (2023). ChatGPT for next generation science learning. *XRDS: Crossroads, The ACM Magazine for Students, 29*(3), 42-46. https://doi.org/10.1145/3589649

Zohny, Hazem; McMillan, John, and King, Mike (2023). Ethics of generative AI. *Journal of Medical Ethics, 49*(2), 79-80. http://dx.doi.org/10.1136/jme-2023-108909