Artificial intelligence versus journalists: The quality of automated news and bias by authorship using a Turing test*

Leonardo Alberto La-Rosa Barrolleta
Universidad de Nebrija
llarosa@nebrija.es

Teresa Sandoval-Martín
Universidad Carlos III de Madrid
msandova@hum.uc3m.es

Abstract

The integration of Artificial Intelligence (AI) in the media results in the publication of thousands of automated news articles in Spanish every day. This study uses a Turing test to compare the quality of news articles written by professional journalists (from Efe) with those produced by natural language generation (NLG) software (from Narrativa). Based on Sundar's dimensions (1999) crucial to news perception – credibility, readability and journalistic expertise – , an internationally validated experimental methodology is employed, exploring a novel topic in Spanish: health information. The experiment deliberately varied real and declared authorships – AI and human journalists – to detect potential biases in assessing authorship credibility. A self-administered questionnaire adapted for online surveys was used (N=222), and gender imbalances were minimized to ensure gender equality

in the sample (N=128). The study reveals that there are no significant differences between news articles generated by the AI and those written by professional journalists. Both types of news are considered equally credible, though some biases are detected in the evaluation of declared authorship: the AI author is perceived as more believable than the human, while the human journalist is perceived as creating a more lively narrative. The study concludes that it is feasible to produce automated news in Spanish without compromising its quality. In the global media landscape, automated systems employing NLG, machine learning and sophisticated databases successfully advance into new domains such as health information.

Keywords: automated journalism; automated news; artificial intelligence; Turing test; COVID-19

Resum. *Intel·ligència artificial enfront de periodistes: qualitat de les notícies automatitzades i biaix per autoria mitjançant una prova de Turing*

Amb la integració de la intel·ligència artificial (IA) en els mitjans cada dia es publiquen milers de notícies automatitzades en espanyol. Aquest estudi compara la qualitat de les notícies escrites per periodistes professionals (d'Efe) amb les realitzades amb programes informàtics de generació de llenguatge natural (NLG) (de narrativa) mitjançant un test de Turing. Basat en dimensions crucials per a la percepció de notícies (Sundar,1999) —credibilitat, llegibilitat i perícia periodística—, s'usa una metodologia experimental comprovada a escala internacional per explorar un tema inèdit en espanyol: informació de salut. L'experiment va variar intencionadament les autories reals i declarades —periodistes humans i IA— per detectar possibles biaixos entorn de la credibilitat de l'autoria. Es va fer servir un qüestionari autoadministrat adaptat per a enquestes en línia (N = 222) i es van minimitzar els desequilibris de gènere per assegurar la paritat de la mostra (N = 128). L'estudi revela que no hi ha diferències significatives entre les notícies generades per IA i les elaborades per periodistes professionals. Tots dos tipus de notícies són considerades igualment creïbles, encara que es van detectar alguns biaixos en l'avaluació de l'autoria declarada: la IA com a autor és percebuda com més creïble que l'humà, mentre que es considera que el periodista humà escriu textos més vívids. L'estudi conclou que és factible produir notícies automatitzades en espanyol sense comprometre'n la qualitat. En el panorama mediàtic global, els sistemes automatitzats que empren GLN, aprenentatge automàtic i sofisticades bases de dades avancen amb èxit cap a nous àmbits, com la informació de salut.

Paraules clau: periodisme automatitzat; notícies automatitzades; intel·ligència artificial; prova de Turing; COVID-19

Resumen. *Inteligencia artificial frente a periodistas: calidad de las noticias automatizadas y sesgo por autoría mediante una prueba de Turing*

Con la integración de la inteligencia artificial (IA) en los medios cada día se publican miles de noticias automatizadas en español. Este estudio compara la calidad de las noticias escritas por periodistas profesionales (de Efe) con las realizadas con programas informáticos de generación de lenguaje natural (NLG) (de narrativa) mediante un test de Turing. Basado en dimensiones cruciales para la percepción de noticias (Sundar,1999) —credibilidad, legibilidad y pericia periodística—, se usa una metodología experimental comprobada a escala internacional para explorar un tema inédito en español: información de salud. El experimento varió intencionadamente las autorías reales y declaradas —periodistas humanos e IA— para detectar posibles sesgos en torno a la credibilidad de la autoría. Se usó un cuestionario autoadministrado adaptado para encuestas en línea (N = 222) y se minimizaron los desequilibrios de género para asegurar la paridad de la muestra (N = 128). El estu-

dio revela que no hay diferencias significativas entre las noticias generadas por IA y las elaboradas por periodistas profesionales. Ambos tipos de noticias son consideradas igualmente creíbles, aunque se detectaron algunos sesgos en la evaluación de la autoría declarada: la IA como autor es percibida como más creíble que el humano, mientras se considera que el periodista humano escribe textos más vívidos. El estudio concluye que es factible producir noticias automatizadas en español sin comprometer su calidad. En el panorama mediático global, los sistemas automatizados que emplean GLN, aprendizaje automático y sofisticadas bases de datos avanzan con éxito hacia nuevos ámbitos, como la información de salud.

Palabras clave: periodismo automatizado; noticias automatizadas; inteligencia artificial; prueba de Turing; COVID-19

1. Introduction

The rise of Artificial Intelligence is a phenomenon that affects all sectors. As we enter the Fourth Industrial Revolution (Schwab, 2016) journalism is no exception – and not without risks and opportunities. In fact, AI changes how information is "obtained, stored, processed, transmitted and consumed" (Túñez-López et al., 2021: 178; Tejedor et al., 2021a). This paper focuses on the quality of automated news from the audience's perspective (Clerwall, 2014; Haim and Graefe, 2017; Moravec et al., 2020; Zheng et al., 2018). Its findings help clarify the question of whether automated news articles achieve the necessary quality to compete with articles written by journalists, setting aside its effects on the profession, the media or even the tensions between them (Moran and Shaikh, 2022).

Advances in the field of information and communication technologies (ICTs), linguistics and natural language generation (NLG) have allowed algorithms to write automated news from stored data in a structured and readable format for machines (Clerwall, 2014; Diakopoulos, 2019; Graefe et al., 2018; Graefe and Bohlken, 2020; Haim and Graefe, 2017; Jia, 2020; Lermann-Henestrosa et al., 2023; Tandoc et al., 2020; Van-Der-Kaa and Krahmer, 2014; Wölker and Powell, 2021).

News organisations have increasingly been using computer-generated news in recent years (Fanta, 2017; Beckett and Yaseen, 2023). North America (*The Associated Press*, *The New York Times*, *Los Angeles Times, The Washington Post, Forbes*), China (*Xinhua News Agency, Southern Metropolis Daily*) and Japan (*Shinano Mainichi Shimbun, Spectee Inc.*) are countries with an early and wider implementation of computer-generated news. In Europe, the United Kingdom and Finland stand out (Linden and Tuulonen, 2019; Túñez-López et al., 2019; Calvo-Rubio and Ufarte-Ruiz, 2020), while Eastern European countries encounter difficulties due to the Slavic origins of their languages (Movarec et al., 2020).

In the field of automated news in Spanish, the company *Narrativa Inteligencia Artificial* (Ufarte-Ruiz and Manfredi, 2019) stands out, serving a multitude of media companies in Spain and other countries – including news agencies. This paper therefore focuses its analysis on information generated

by the company, as its services rely on NLG and machine learning and it reaches a significant proportion of the Spanish-speaking population through the media outlets that use them.

Automated news on different topics has been analysed before in experiments that based their dimensions linked to quality – credibility, readability and expertise – on Sundar's approach (1999) and his MAIN model: "Modality (M), Agency (A), Interactivity (I) and Navigability (N)" (Sundar, 2008: 78). Most of the topic stimuli were around finance, sports, entertainment and politics, though recently topics more related to complex texts have been added, such as scientific information around conferences and talks, biodiversity and technology (Table 1). However, health information was not analysed in these experimental designs, and specifically news generated during the COVID-19 pandemic lockdown.

Table 1. Topic stimuli in previous experimental designs

Topic stimuli	Authors
Sports	Clerwall (2014)
Politics	Jia and Jhonson (2021), Waddell (2019)
Breaking news: earthquake alerts	Tandoc et al. (2020)
Finance and sports	Van-Der-Kaa and Krahmer (2014), Wölker and Powell (2021), Graefe et al. (2018)
Finance, entertainment and sports	Haim and Graefe (2017)
Finance, sports and politics	Wu (2019)
Finance, sports, entertainment (civic news) and Scientific information (conference news)	Jia (2020)
Scientific information: biodiversity, technology	Lermann-Henestrosa et al. (2023)

Source: Author's own.

As Knight News Innovation fellow Danzon-Chambaud notes:

> When the virus spread globally at the beginning of 2020, governments and health authorities made accessible a considerable amount of open-source data, generally available through structured datasets or APIs. These statistics contained critical information such as the number of deaths and patients in intensive care units as well as 7-day incidence rates. (Danzon-Chambaud, 2023: 98)

The abundance of ever-expanding structured data made the COVID-19 pandemic an opportune time to automate tables, news and interactive graphics. Moreover, the media assumed a "pivotal role in disseminating information to the public regarding the virus spread, especially during the most critical phases of lockdown" (Tejedor et al., 2021b: 252).

Several media organisations attempted to leverage the advantages offered by this open-data to automate structure data from COVID-19 that could result in predictable story frames. In a study that compiled the experience of nine media organisations, Danzon-Chambaud found that working with

external databases and AI signified a move towards "new forms of newsroom workflows that involved contributing to the creation of automated news or working directly with them" (Danzon-Chambaud, 2021).

In this context, *Narrativa* launched its purportedly non-profit COVID-19 Monitoring Project at the beginning of the pandemic. According to their website, the project aimed to gather "information from different data sources to provide comprehensive data on the new coronavirus, SARS-CoV-2,". At the same time, the largest Spanish news agency, *Efe*, which annually disseminates approximately three million news items (SEPI, 2020), formulated a strategic plan to mitigate the economic repercussions of the pandemic. This plan involved technological advances and digitalisation, and enabled the agency to maintain its leadership in Spanish-speaking countries (SEPI, 2021).

These reasons made *Efe* and *Narrativa* the perfect fit for an experiment that aims to analyse how news written by AI is perceived by media users in Spanish, a language spoken by 595 million people worldwide (Instituto Cervantes, 2022). Although previous studies have analysed how journalists and users perceive automated news in Spanish – as we will discuss in the following section – , no study had ever been conducted using widely tested experimental methods validated internationally (Sandoval-Martín and La-Rosa, 2023). This paper fills that gap. The reliability of the experimental methodology employed adds value to the results of the study.

The growing platformisation of the media industry and the economic challenges faced by news organisations increase the need to produce brief, quick and cost-effective news items, potentially leading to an increase in the production of automated news worldwide. This study aims to contribute to a better understanding of how to analyse and measure the quality of automated news in Spanish using widely accepted experimental methods. It also explores the prospect of diversifying topics to encompass health-related information.

2. Literature review

Ever since the English mathematician and widely-acknowledged father of computing Alan Turing started to speak about his automatic-intelligent machine in 1936 and created the well-known Turing test, researchers from various disciplines have been intrigued by the possibility that computers could exhibit intelligence, albeit for specifics periods (Parra and Torrens, 2018). In the field of journalism, robotic storytelling has been in existence for over 40 years (Meehan, 1977). Journalists have used computers and computing programmes to investigate facts or analyse data at least since the Seventies (Meyer, 1973). However, it was not until the era of big data – within a distinctive technological and sociocultural context (Sandoval-Martín and La-Rosa, 2018) – that its use expanded, coinciding with the rise of artificial intelligence at the end of the last decade. In this new context and as an extension of 'data journalism', the media industry has begun to innovate and apply AI in the production of news. In early academic articles, this development was

given various names: *robot journalism*, *algorithmic journalism*, *machine-written news* or *computational journalism* (Lee and Kim, 1998; Matsumoto et al., 2007; Latar and Nordfors, 2009; Latar, 2014; Clerwall, 2014; Domingos, 2015; Rutkin, 2014; Dörr, 2016). In the field of the sociology of communication *automated journalism* (AJ) was preferred (Carlson, 2015; Lecompte, 2015; Napoli, 2014), and this term has become increasingly widespread.

For AI, this study adopts the widely agreed definition issued by the European Commission's High-Level Expert Group on AI (European Commission, 2018). This definition encompasses software and hardware systems designed by humans for complex purposes that perceive and interpret their environment through the collection of data, whether structured or not. These systems engage in reasoning using knowledge, and process the information derived from it, ultimately deciding the best course of action to achieve a given objective. Furthermore, AI can interact with its environment and can modify its behaviour based on the reactions that have occurred in response to its previous actions.

The quality of journalistic narratives has traditionally been a complex and often overlooked issue. This complexity is further heightened in the realm of AI systems. A systematic review of the scientific literature on research into the quality of artificial journalism reveals that it has gained significant attention since the middle of the last decade (Sandoval-Martín and La-Rosa, 2023). Studies focused on how algorithms can autonomously generate news (Carlson, 2015); the extent of automation in newsrooms (Lecompte, 2015); the processes involved in content production and consumption (Napoli, 2014); and notably, how news written by robots is perceived (Clerwall, 2014; Haim and Graefe, 2017; Moravec et al., 2020; Zheng et. al, 2018).

Additionally, various studies have explored confirmation bias, i.e., evaluator bias when aware of (human-machine) authorship (Jia and Johnson, 2021; Jung et al., 2017; Lermann-Henestrosa et al., 2023; Waddell, 2019; Wölker and Powell, 2021; Tandoc et al., 2020). Furthermore, other works employing different approaches have examined similarities and differences between automated and journalist-written texts (Murcia Verdú et al., 2022) as well as patterns in their structures (Ufarte-Ruiz and Manfredi, 2019), among others, in an attempt to elucidate the quality of automated news. Thus, the results of previous studies in other languages (Haim and Graefe, 2017; Wölker and Powell, 2021; Wu, 2019; Moravec et al., 2020; Jia, 2020; Jia and Johnson, 2021) agree that "the automated news quality is generally perceived as optimal, although with some limitations such as the impossibility of adding context, different points of view and interpretation" (Sandoval-Martín and La-Rosa, 2023: 119). Most researchers have been inspired by Shyam Sundar's 1999 study on the perception of print and online news and have applied an experimental design (Sandoval-Martín and La-Rosa, 2023).

In most experiments evaluating the quality of automated news in other languages, no significant differences were found between news articles generated by AI and those written by professional journalists, although small biases in authorship were observed (Table 2).

Table 2. Findings of experimental designs that evaluate the quality of automated news based on Sundar's approach

Authorship	Findings on perception of the message and authorship bias
Clerwall (2014)	No differences in the quality of news written by AI and human journalists in credibility and readability items.
Van-Der-Kaa and Krahmer (2014)	No differences in the quality of news written by AI and human journalists.
Haim and Graefe (2017)	Little difference in the quality of news written by AI and human journalists.
Graefe et al. (2018)	Differences in terms of perceived credibility and expertise tend to be small, favouring evaluation of automated news in the message. Little bias was found in authorship: human-written texts were better rated in authorship credibility items.
Waddell (2018)	Small biases in authorship: human-written texts better perceived in credibility.
Waddell (2019)	Little bias in authorship: declared AI authorship decreased perception of credibility of news.
Wu (2019)	Automated news perceived as more credible (message), but significant differences – not biases – were found in perception of authorship depending on the topic.
Jia (2020)	No significant difference in terms of credibility of the text. Significant differences in human-written news in readability and expertise items.
Tandoc et al. (2020)	No significant differences perceived in message and authorship credibility between algorithm, human, and mixed (human-AI).
Jia and Johnson (2021)	No significant difference in terms of credibility of the text, but small differences were detected toward human-journalist authorship over automated news – not biases. For both human and algorithm authorship, attitude-consistent news is rated as more credible than attitude-challenging news.
Wölker and Powell (2021)	Perceived credibility of AI-written news, human-written news and mixed (human and AI) may be assumed equal. No bias detected towards AI-written news in authorship.
Lermann-Henestrosa et al. (2023)	No differences in perceived credibility and trustworthiness between AI-written and human-written texts. No authorship biases detected.

Source: Author's own.

The present study provides knowledge in this field by revealing that in the Spanish language the audience do not perceive significant differences between news articles generated by AI and those written by professional journalists. Both types of news are considered equally credible, although some biases were detected in the evaluation of declared authorship, as discussed in the Results section of this paper.

This research aligns with the field of Science, Technology and Society (STS) studies, which addresses issues relating to scientific and technological practices affecting everyday life. Within this context, the emergence of automated news raises crucial questions, such as challenging the traditional role of human journalists as sole content creators in journalistic narratives.

3. Hypothesis and objectives

The general purpose of this study is to assess whether Spanish-speaking audiences perceive automated news to be of similar quality to human-written news. The first hypothesis (H1) posits that Spanish audiences perceive automated news to be of comparable quality to human-written news, particularly in the context of health-related information. In other words, AI seems to provide credible and high-quality information about health, even in the case of automated news in Spanish. A subsequent hypothesis (H2) explores biases relating to authorship: there are positive biases in people's evaluation of news articles created by machines.

The primary objective is to compare audience perception of news articles written in Spanish by either a robot or a human journalist with real or declared authorship, taking into account three dimensions: credibility, readability and journalistic expertise. A specific objective, linked to this primary one, is to determine the credibility and readability of the message in each case and level – implicit in quality.

The second objective aims to determine authorship credibility in each case, so it will be possible to identify potential biases in people's evaluation of news articles created by machines versus those written by humans. This involves investigating whether informing evaluators about the authorship of the news article influences their response regarding authorship credibility.

The third general objective is to assess journalistic expertise. A subsequent specific objective involves identifying positive or negative biases for journalists and machines in the authorship variable.

4. Methodology and experimental design

Given the lack of transparency regarding algorithms used for automated news generation, a tool involving reverse engineering is needed. A methodological instrument to measure the quality of this kind of news content with journalistic parameters is a first step to building up an automated reverse-engineering tool. This study employs a qualitative methodological approach to analyse how users perceive the quality of automated news compared to news made by humans, creating an instrument to achieve the first step of an automated reverse-engineering tool.

The instrument is based on previous experimental methodologies in Sweden (Clerwall, 2014), the Netherlands (Van-Der-Kaa and Krahmer, 2014), Germany (Haim and Graefe, 2017; Graefe et al., 2018; Graefe and Bohlken, 2020) and several Asian countries (Wu, 2019; Jia, 2020; Tandoc et al., 2020). A methodological design involving a Turing test to evaluate automated news-writing in natural language for Spanish news consumers is presented, to exhibit the skills of automated news software in performing like a human journalist according to Turing's Imitation Game (Turing, 1950).

The experiment involves presenting respondents with two news articles based on the same information extracted from COVID-19 databases, one written automatically by *Narrativa* and one by a human journalist from *Efe* (Appendix A). The study uses news created from large health databases, such as the official open data generated during the pandemic by the Spanish Ministry of Health and the health councils of 17 Autonomous Communities (Guisado-Clavero et al., 2022).

After reading each text, respondents answer questions relating to news quality based on specific survey objectives and theoretical and methodological foundations presented in the following sections. This approach allows the quality of news written by a computer to be analysed and compared to articles on the same news event written by journalists. To eliminate biases and detect potential evaluator biases in assessing authorship credibility, two news articles with falsified authorship were included. One article written by a journalist was falsely attributed to a computer, and vice versa.

The Turing test used a self-administered questionnaire adapted for online surveys and experimental methodologies. It draws on psychological dimensions relevant to news perception: credibility, readability and journalistic expertise (Sundar, 1999). Respondents evaluated news quality based on these three dimensions. The original questionnaire can be found in the supplementary material (Appendix B).

To evaluate credibility, respondents were asked dichotomous questions for six message credibility items and four readability items. Message credibility was assessed based on Meyer's (1988) and Flanagin and Metzger's (2000) criteria, adapted to evaluate the credibility of automated news (Wölker and Powell, 2021). These items assess whether the message is: 1) credible or noncredible; 2) accurate or inaccurate; 3) impartial or biased; 4) balanced or unbalanced; 5) reliable or unreliable; and 6) complete or incomplete. Readability items gauge whether the message is: 1) entertaining or boring; 2) lively or not lively; 3) interesting or uninteresting; and 4) well-written or poorly written.

The second part of the evaluation focused on message credibility, using the same six items mentioned above. Additionally, it assessed journalistic expertise using four items, evaluating whether the message was coherent, concise, descriptive and comprehensive. Respondents evaluated these 10 items on a 5-point Likert-scale, where 1 represents the most negative evaluation and 5 the most positive.

4.1. Definition and operationalisation of variables

Each of the desirable items is represented in Table 3 along with its respective dimensions. The table also explains what is being assessed with each item. This was available during the assessment process, to resolve any doubts for respondents. Since there are items that are useful for assessing both the message and the authorship, their explanations have been grouped according to the object of the assessment.

Table 3. Quality dimensions of news and items to assess

Dimensions	Message	Authorship
Credibility	· Believable · Accurate · Unbiased · Complete · Reliable · Fair	· Believable · Accurate · Unbiased · Complete · Reliable · Fair
Readability	· Entertaining · Lively · Interesting · Well written	-
Expertise	-	· Coherent · Comprehensive · Concise · Descriptive

Item	Message	Authorship
Believable	Measures whether the message is believable.	Measures the credibility of the authorship.
Accurate	Measures whether the message comes as close as possible to what really happened.	Measures whether the authorship presents verifiable facts.
Unbiased	Measures whether the message does not lean toward any position.	Measures whether the authorship does not take sides in the narration of facts.
Complete	Measures whether the message does not lack any information that helps to explain things better.	Measures whether the authorship has not omitted any relevant information.
Reliable	Measures whether the message conveys confidence in the information presented.	Measures whether the authorship is worthy of confidence or seems to be correct.
Fair	Measures whether the message has an informative balance between positions.	Measures whether the authorship has been fair when presenting all sides of the story.
Entertaining	Measures whether the message has been written in an entertaining manner.	-
Lively	Measures whether the message is capable of evoking great force and clarity.	
Interesting	Measures whether the message contains information that might interest the reader.	
Well-written	Measures how well written the message is, if it has spelling, grammatical or drafting errors.	
Coherent	-	Measures whether the authorship relates the narrated events logically.
Comprehensive		Measures whether the authorship narrates the events in an understandable manner.
Concise		Measures the authorship's degree of expressiveness (few and adequate words).
Descriptive		Measures the authorship's capacity to describe events in great detail.

Source: Author's own.

4.2. Respondents and conditions assessed

The experiment was administered via an online survey (N=222) in May 2020, in which each respondent was assigned a text for reading with one of the following conditions (Table 4).

Table 4. Authorship conditions of news content

Cases	Real authorship	Declared Authorship	Respondents	Code Name
Case 1	Human journalist	Human journalist	N=37	Real Jour. * Dec. Jour
Case 2	Human journalist	AI	N=42	Real Jour. * Dec. AI
Case 3	AI	Human journalist	N=50	Real AI * Dec. Jour
Case 4	AI	AI	N=93	Real AI * Dec. AI

Source: Author's own.

The news article that respondents in cases 1 and 2 read was written by journalists from the *Efe* news agency on 4th May 2020, while the news article that respondents in cases 3 and 4 read was written by the automated software of *Narrativa* on the same date. Both news articles were about COVID-19 and had very similar headlines and numbers of characters (Appendix A). The two news articles gave daily figures of individuals affected by the virus, accessible on the internet – on the websites of both the technology company and the news agency. Only the text was displayed, and the design was standardised to avoid interference with perception. Tables were removed in the case of the automated news, as they contained additional data and information already present in the text.

After excluding incomplete surveys, the sample was reduced by excluding assessments completed by people over 40 years of age and an excess number of men, at random, to obtain an equal sample (N=128), with 32 respondents for each case.

A total of 51.56% of assessments were completed by women, while two subjects preferred not to declare their gender. The mean age was 21.68 years old – standard deviation (SD)=5.846 – and the nationality of 90% of the sample was Spanish. Only 15% of respondents had had no education related to journalism at all: out of 100 who had finished high school, 39 were journalism students and 57 were students of combined degree programmes including journalism. In this way, value has been added to the experiment by incorporating a high percentage of qualified respondents.

4.3. Procedure

Over the course of a week during term time, links to the online surveys were provided to the staff and students of the undergraduate degree and combined degree programmes in Journalism at Universidad Carlos III de Madrid. They in turn shared the links with colleagues and acquaintances. The links were provided randomly, and each one corresponded to one of the conditions of the experimental design.

Respondents provided demographic information such as age and level of education, among others. They then proceeded to read the assigned text in order to evaluate it. They answered the questionnaire regarding the credibility and readability of the message first, followed by the questionnaire relating to the credibility and journalistic expertise of the authorship.

In each of the four conditions presented, Kolmogorov-Smirnov and Shapiro-Wilk normality tests were conducted for each of the 20 variables, using the statistical software SPSS. None of the variables had the necessary normality for the application of parametric tests. Therefore, a sampling simulation or *bootstrap* of 1,000 samples in each variable was made, a versatile method to estimate the sampling distribution of estimated parameters (Efron, 1982) and thus find approximate standard errors of the variables for comparison purposes. Initial descriptive results of the demographic variables were also made in this way.

Using this method, descriptive statistics were extracted, and three one-way analysis of variance (ANOVA) tests were conducted for each of the dependent variables with the following independent variables: 1) Real Authorship; 2) Declared Authorship; and 3) Authorships (Real Authorship*-Declared Authorship). In the cases of Real Authorship and Declared Authorship, the design was A=2, while in the case of Authorships, it was A=4.

For the ANOVA analysis where the factor was Authorships, a *post hoc* comparison in SPSS was performed to determine which pairs of means exhibited significant differences. Tukey's Honestly Significant Difference (HSD) test and the Games-Howell test were conducted for all variables, interpreting the results based on the homogeneity of variance provided by the Levene's statistic.

Lastly, our criterion for selecting significant differences from the three ANOVA analyses conducted was based on our interest in detecting biases among the evaluators when assessing authorships. In this regard, significant differences in declared authorships took precedence over real and mixed authorships. In cases where no significant differences were found in declared and real authorships, consideration was given to those involving authorships that, after a post hoc comparison, did not imply variance homogeneity according to the Levene's statistic.

5. Results

In the analysis of responses from the first part of the questionnaire, reading habits regarding both general topics and COVID-19 information did not show significant differences across the variables studied. No gender or education-related disparities were observed. However, variations emerged based on specialisation: respondents without a journalism background tended to rate the authorship balance more positively, scoring it at 3.74, compared to 3.31 and 2.88 for the two groups with a journalism education. The most important findings in the second part of the questionnaire will be explained in the following paragraphs.
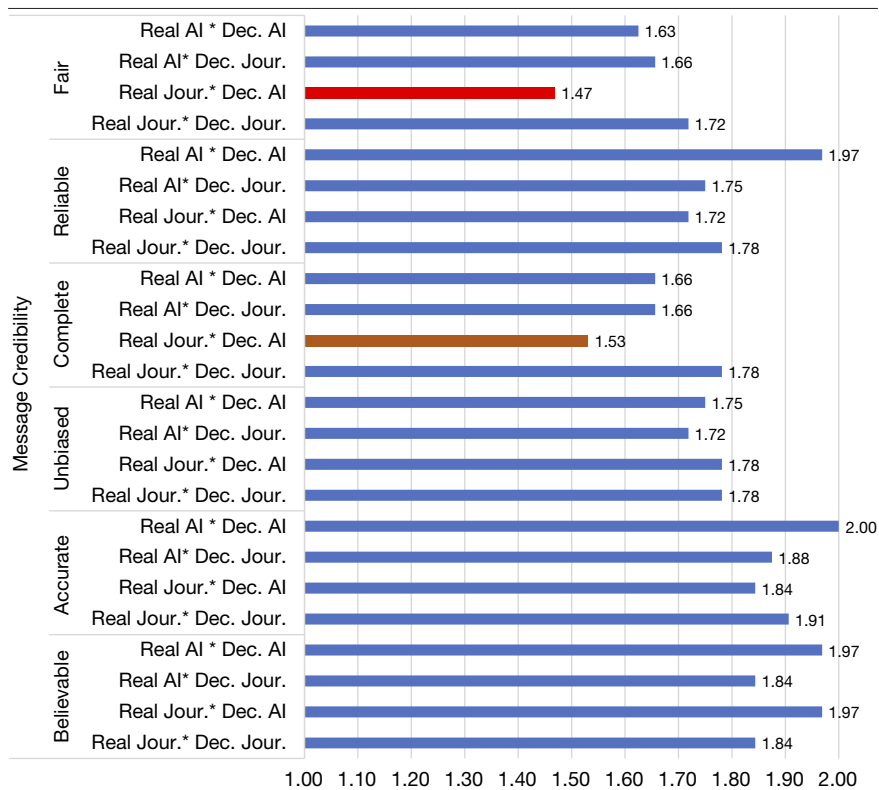
5.1. Credibility and readability of the message

In general, the articles written by human journalists and those written by AI were perceived as credible, taking into account the following regarding credibility:

Apart for the variables 'Complete' and 'Fair', when comparing cases 1 and 4, AI was rated better than journalists. Nevertheless, the difference between the two is small. If we restrict ourselves to a comparison of these two cases, the differences between the variables 'Reliable' (AI=1.97 and Jour.=1.78) and 'Believable' (AI=1.97 and Jour.=1.84) are much more significant, even though the lowest means of these variables are still high (Figure 1).

In Case 2, the variable 'Fair' does not exceed the neutral point of 1.50. Nevertheless, the same text signed by a journalist received the highest rating of all the cases in the same variable (1.72), alone exceeding the two cases relating to the text written by AI. This means that the message of the text written by journalists was not perceived as less balanced or fair than the one written by AI.

**Figure 1.** Means of the variables of the message credibility in each case

Source: Author's own.

If we look at significant differences in assessments of the credibility and readability of the message, these only appear in the declared authorship of the variables 'Believable' and 'Lively': the respondents perceive a text as more believable when it is signed by AI (Figure 1) and perceive it to be more lively when it is signed by a journalist (Figure 2).

It is worth noting that in the message variables for the dimension 'Readability', none of the means of the cases reached the neutral value of 1.5 in 'Entertaining' and 'Lively' (Figure 2).

**Figure 2.** Means of the variables of the message readability in each case
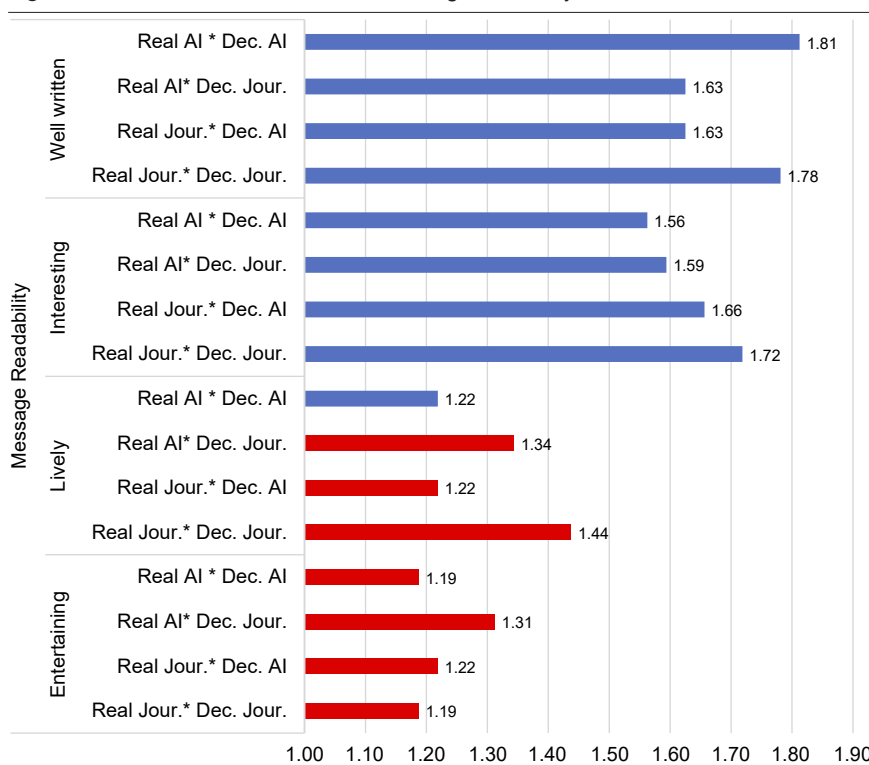
Source: Author's own.

5.2. Authorship credibility

The bias shown in the variable 'Believable' seems to be confirmed by the general result of the items relating to Authorship in the dimension 'Credibility' (Figure 3), since AI was rated more highly than real journalists. Nonetheless, the only significant difference worth taking into account in this dimension was in the declared authorship of the Authorship variable 'unbiased', where a positive bias was found in the texts signed by AI compared to those signed by journalists.

Figure 3. Means of the variables of the authorship credibility in each case
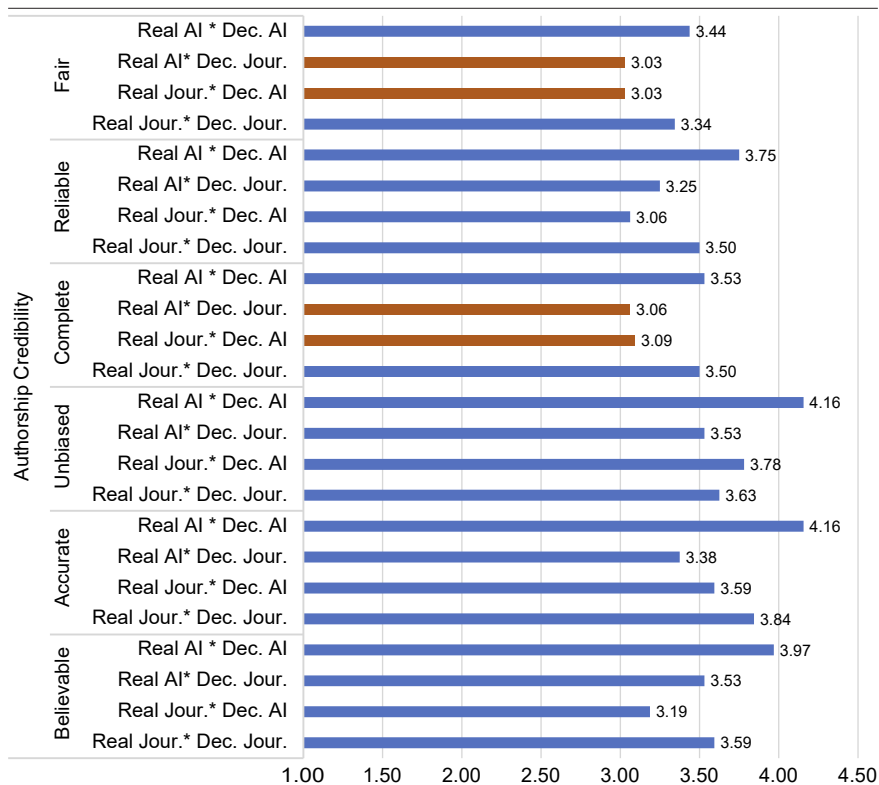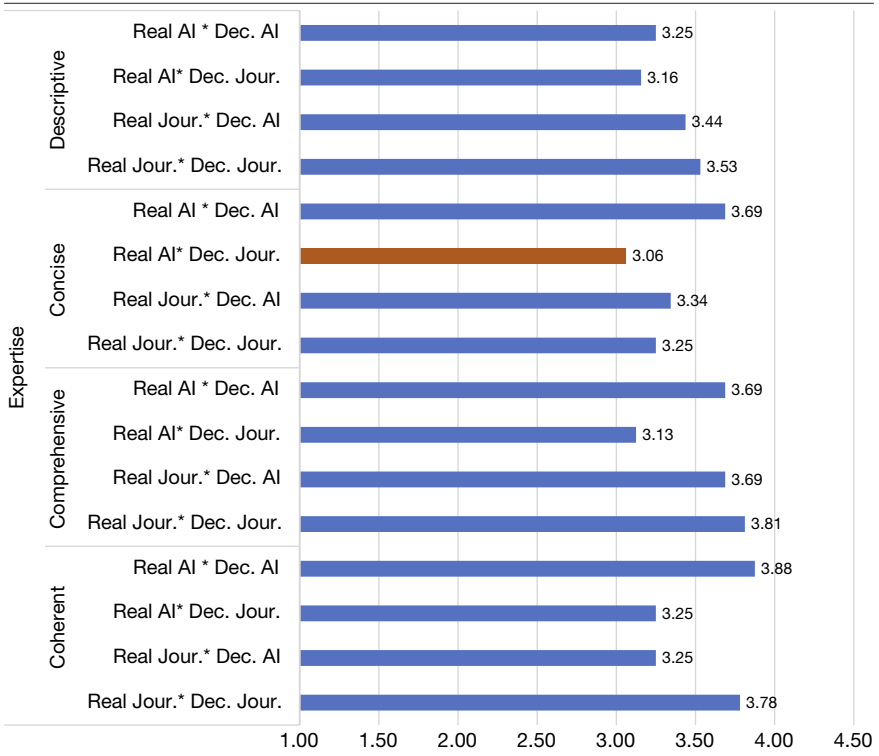

Source: Author's own.

It is also noticeable how in the general results some variables slightly miss the neutral value: 'Complete', 'Reliable' and 'Fair'.

5.3. Journalistic expertise

In the general comparison of the cases in the dimension 'Expertise', journalists are more descriptive and comprehensible than AI, while the latter is more concise (Figure 4). The only significant difference appeared in the Authorship variable 'Comprehensive', where it can be observed that journalists were rated more highly by respondents, indicating a bias by respondents when rating comprehensibility.

Figure 4. Means of the variables of the authorship expertise in each case

Source: Author's own.

6. Discussion and Conclusions

Overall, the content written by human journalists and AI are perceived as credible in both message and authorship (H1). This is a common result in previous studies, and is due to the illusion that articles written by algorithms are more objective than those written by journalists. This has been corroborated empirically by the detection of a bias that leads to texts written by AI being evaluated more positively in terms of the credibility of the text and the impartiality of the authorship than written by journalists (H2). In this regard, the research results coincide with those of Wölker and Powell (2021) in their study on automated sports news.

This indicates the enormous danger in perceiving algorithms to be impartial, since it confirms that the general public does not take into account the fact that algorithms are written with the biases of their human creators.

Both AI and journalists are perceived to have journalistic expertise. However, we cannot be sure that their authorships were perceived as readable, since they were not considered to be entertaining or lively. The latter could be

related to the topic, since delivering figures on the state of the pandemic did not leave any room to write an entertaining text, with great force and clarity. Nevertheless, all cases greatly surpassed the other variables, 'Interesting' and 'Well-written', in the message readability dimension (H1). Therefore, in spite of not entertaining and not being written lively, the topic concerned and interested everyone, and the texts were considered to be well written by both journalists and AI. Based on these results, the media and journalists should consider producing writing that is perceived to be less boring.

The fact that the automated news pieces were not considered to be entertaining or lively in the case of the chosen topic raises the possibility of continuing the experiment with other topics (finance, sport, the weather) and making comparisons. However, the expected results would be similar to those obtained in this experiment if we take into account a fact that has been highlighted by authors of previous studies (Graefe et al., 2018; Wölker and Powell, 2021): human journalists who are asked to write numerous articles at great speed tend to list facts and figures repetitively, like algorithms, on routine topics.

Curiously, in the variables of the 'Readability' dimension, in which all the cases exceeded the neutral value, the differences between means are not very great, so the statements added to the text written by a human did not really make this topic seem less of a routine task. However, if we take into account that the only significant difference in this dimension appeared precisely in how the message was perceived as lively, we can infer that the fact that the text written by a journalist included textual quotes and the one from the AI did not is closely related to the fact that texts written by journalists are perceived as more lively than those written by AI. All of this in spite of having found a bias in which people tend to rate texts written by journalists as more lively.

Given the results produced in this study, we can confirm that in the Spanish context, algorithms can coexist with journalism professionals in making automated news. This form of narrative is integrated into the public's media consumption without their conscious knowledge.

The author of *Automating the News* (Diakopoulos, 2019) uses the term 'hybrid journalism' to refer to a new environment in which complex communication, expert thinking and ethical judgment of human journalists will still have a lot to add. It is precisely within this context that examining the quality of automated journalism becomes crucial, especially following the recent opening to the public of large language models (LLMs). US-based ChatGPT and China-based Ernie – released in November 2022 and March 2023 respectively – are LLMs, advanced language models with massive parameter sizes and exceptional learning capabilities (Chang et al., 2023). In other words, "the most prominent form of generative AIs" (Arguedas and Simon, 2023). This will be the focus of our coming research projects.

When a LLM scrapes gigantic swathes of information from the internet to train the machine, there is not enough transparency within this process to confirm sources. Even less large NLG tools used in specific fields such as journalism might have the same problem, making the study of audience perceptions

of automated content an area that the academic community needs to cover. If automated news can be perceived to be as good as a story written by a journalist, the transparency issues highlighted above will become even more critical.

Bibliographical references

ARGUEDAS, A.R. and SIMON, F.M. (2023). *Automating Democracy: Generative AI, journalism, and the future of democracy.* Oxford: Oxford Internet Institute.
<https://doi.org/10.5287/ora-e262xv7no>

BECKETT C. and YASEEN, M. (2023). *Generating Change. A global survey of what news organisations are doing with AI.* London: LSE. Retrieved from <https://www.journalismai.info/research/2023-generating-change>

CALVO-RUBIO, L.-M. and UFARTE-RUIZ, M.-J. (2020). "Percepción de docentes universitarios, estudiantes, responsables de innovación y periodistas sobre el uso de inteligencia artificial en periodismo". *Profesional de la información*, 29 (1).
<https://doi.org/10.3145/epi.2020.ene.09>

CARLSON, M. (2015). "The Robotic Reporter". *Digital Journalism*, 3 (3), 416-431.
<https://doi.org/10.1080/21670811.2014.976412>

CHANG, Y., WANG, X., WANG, J., WU, Y., YANG, L., ZHU, K., CHEN, H., YI, X., WANG, C., WANG, Y., YE, W., ZHANG, Y., CHANG, Y., YU, P.S., YANG, Q. and XIE, X. (2023). "A Survey on Evaluation of Large Language Models". *ArXiv preprint arXiv:2307.03109v7*, 1-26.
<https://doi.org/10.48550/arXiv.2307.03109>

CLERWALL, C. (2014). "Enter the Robot Journalist. Users' perceptions of automated content". *Journalism Practice*, 8 (5), 519-531.
<https://doi.org/10.1080/17512786.2014.883116>

EUROPEAN COMMISSION (2018). *A definition of Artificial Intelligence.* Brussels: European Commission. Retrieved from <https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf>

DANZON-CHAMBAUD, S. (2021). "Tow Report. Covering COVID-19 with automated news". *Columbia Journalism Review.* Retrieved from <https://www.cjr.org/tow_center_reports/covering-covid-automated-news.php>

— (2023). *Automated news in practice: Changing the journalistic doxa during COVID-19, at the BBC and across media organizations* [Doctoral thesis, School of Communications Dublin City University]. Doras. Retrieved from <https://doras.dcu.ie/28004/1/DEPOSIT_JANUARY_2023_Danzon-Chambaud_Automated_News_Practice_PhD_Thesis.pdf>

DIAKOPOULOS, N. (2019). *Automating the News.* MA: Harvard University Press.

DOMINGOS, P. (2015). *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World.* St Ives: Allen Lane.

DÖRR, K.N. (2016). "Mapping the Field of Algorithmic Journalism". *Digital Journalism*, 4 (6), 700-722.
<https://doi.org/10.1080/21670811.2015.1096748>

EFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* California: Standford University.
<https://doi.org/10.1137/1.9781611970319>

FANTA, A. (2017). *Putting Europe's Robots on the Map: Automated journalism in news agencies.* Oxford: Reuters Institute for the Study of Journalism.

FLANAGIN, A.J. and METZGER, M.J. (2000). "Perceptions of Internet Information Credibility". *Journalism & Mass Communication Quarterly*, 77 (3), 515-540.
<https://doi.org/10.1177/107769900007700304>

GRAEFE, A. and BOHLKEN, N. (2020). "Automated journalism: A meta-analysis of readers' perceptions of human-written in comparison to automated news". *Media and Communication*, 8 (3), 50-59.
<http://dx.doi.org/10.17645/mac.v8i3.3019>

GRAEFE, A., HAIM, M., HAARMANN, B. and BROSIUS, H.-B. (2018). "Reader's perception of computer-generated news: credibility, expertise, and readability". *Journalism*, 19 (5), 595-610.
<https://doi.org/10.1177/1464884916641269>

GUISADO-CLAVERO M., ASTIER-PEÑA, M.P., GÓMEZ-BRAVO, R. and ARES-BLANCO, S. (2022). "Open data for monitoring COVID-19 in Spain: descriptive study". *Enfermedades Infecciosas y Microbiología Clínica*, 42 (2), 80-87.
<https://doi.org/10.1016/j.eimc.2022.10.002>

HAIM, M. and GRAEFE, A. (2017). "Automated News: Better than expected?". *Digital Journalism*, 5 (8), 1044-1059.
<https://doi.org/10.1080/21670811.2017.1345643>

INSTITUTO CERVANTES (2022). *El español: una lengua viva. Informe 2022*. Retrieved from <https://cvc.cervantes.es/lengua/espanol_lengua_viva/pdf/espanol_lengua_viva_2022.pdf>

JIA, C. (2020). "Chinese automated journalism: A comparison between expectations and perceived quality". *International Journal of Communication*, 14, 2611-2632. Retrieved from <https://ijoc.org/index.php/ijoc/article/download/13334/3083>

JIA, C. and JOHNSON, T. (2021). "Source Credibility Matters: Does Automated Journalism Inspire Selective Exposure?". *International Journal of Communication*, 15, 3760-3781.

JUNG, J., SONG, H., KIM, Y., IM, H. and OH, S. (2017). "Intrusion of software robots into journalism: the public's and journalists' perceptions of news written by algorithms and human journalists". *Computers in Human Behavior*, 71, 291-298.
<https://doi.org/10.1016/j.chb.2017.02.022>

LATAR, N. (2014). "The robot journalist in the age of social physics: the end of human journalism?". In: G. EINAV (Ed.). *The New World of Transitioned Media*. New York: Springer International Publishing, 65-80.
<https://doi.org/10.1007/978-3-319-09009-2_6>

LATAR, N. and NORDFORS, D. (2009). "Digital Identities and Journalism Content: How Artificial Intelligence and Journalism May Co-Develop and Why Society Should Care". *Innovation Journalism*, 6 (7).

LECOMPTE, C. (2015). "Automation in the Newsroom". *Nieman Reports*, 69 (3), 32-45. Retrieved from <http://niemanreports.org/wp-content/uploads/2015/08/NRsummer2015.pdf>

LEE, S.M. and KIM, T.Y. (1998). "A news on demand service system based on robot agent". In: *International Conference on Parallel and Distributed Systems Proceedings*, 528-532.
<https://doi.org/10.1109/ICPADS.1998.741128>

LERMANN-HENESTROSA, L., GREVING, H., KIMMERLE, J. (2023). "Automated journalism: The effects of AI authorship and evaluative information on the perception of a science journalism article". *Computers in Human Behavior*, 138.
<https://doi.org/10.1016/j.chb.2022.107445>

Linden, C. and Tuulonen, H. (2019). *News Automation. The rewards, risks and realities of 'machine journalism'.* WAN-IFRA. Retrieved from <https://wan-ifra.org/insight/report-news-automation-the-rewards-risks-and-realities-of-machine-journalism/>

Matsumoto, R., Nakayama, H., Harada, T. and Kuniyoshi, Y. (2007). "Journalist Robot: Robot System Making News Articles from Real World". In: *2007 IEEE Intl conf on robotics and automation*, 1234-1241.
<https://doi.org/10.1109/IROS.2007.4399598>

Meehan, J.R. (1977). "Tale-Spin, an Interactive Programme That Writes Stories". In: *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, 91-98. Retrieved from <https://www.ijcai.org/Proceedings/77-1/Papers/013.pdf>

Meyer, P. (1973). *Precision journalism: A reporter's introduction to social science methods.* Bloomington: Indiana University Press.

— (1988). "Defining and measuring credibility of newspapers: Developing an index". *Journalism & Mass Communication Quarterly*, 65 (3), 567-574.
<https://doi.org/10.1177/107769908806500301>

Moran, R.E. and Shaikh, S.J. (2022). "Robots in the news and newsrooms: Unpacking meta-journalistic discourse on the use of artificial intelligence in journalism". *Digital Journalism*, 10 (10), 1756-1774.
<https://doi.org/10.1080/21670811.2022.2085129>

Moravec, V., Macková, V., Sido, J. and Ekštein, K. (2020). "The Robotic Reporter in The Czech News Agency: Automated Journalism and Augmentation in the Newsroom". *Communication Today*, 11 (1), 37-52. Retrieved from <https://communicationtoday.sk/wp-content/uploads/03_MORAVEC-et-al_CT-1-2020.pdf>

Murcia Verdú, F. J., Ramos Antón, R. and Calvo Rubio, L. M. (2022). "Comparative analysis of the sports chronicle quality written by artificial intelligence and journalists". *Revista Latina de Comunicación Social*, 80, 91-111.
<https://doi.org/10.4185/RLCS-2022-1553>

Napoli, P.M. (2014). "Automated Media: An Institutional Theory Perspective on Algorithmic Media Production and Consumption". *Communication Theory*, 24 (3), 340-360.
<https://doi.org/10.1111/comt.12039>

Parra, S. and Torrens, M. (2018). *La inteligencia artificial: el camino hacia la ultrainteligencia*. Barcelona: National Geographic.

Rutkin, A. (2014). "Rise of Robot Reporters: When Software Writes the News". *New Scientist*, 221. Retrieved from <https://www.newscientist.com/article/dn25273-rise-of-robot-reporters-when-software-writes-the-news/>

Sandoval-Martín, T. and La-Rosa, L. (2018). "Big Data as a differentiating sociocultural element of data journalism: the perception of data journalists and experts". *Communication & Society*, 31 (4), 193-209.
<https://doi.org/10.15581/003.31.4.193-209>

— (2023). "Investigación sobre la calidad de las noticias automatizadas en la producción científica internacional: metodologías y resultados". *Cuadernos.info*, 55, 114-136.
<https://doi.org/10.7764/cdi.55.54705>

Schwab, K. (2016). *La Cuarta Revolución Industrial* (The Fourth Industrial Revolution). Barcelona: Debate.

SEPI (2020). *Informe Anual 2020*. Sociedad Estatal de Participaciones Industriales. Retrieved from <https://www.sepi.es/sites/default/files/2022-03/informe_anual_2020_grupo_sepi_0.pdf>

— (2021). *Informe Anual 2021*. Sociedad Estatal de Participaciones Industriales. Retrieved from <https://www.sepi.es/sites/default/files/2022-12/informe_anual_2021_grupo_sepi.pdf>

SUNDAR, S. (1999). "Exploring receivers' criteria for perception of print and online news". *Journalism & Mass Communication Quarterly*, 76 (2), 373-386.
<http://dx.doi.org/10.1177/107769909907600213>

— (2008). "The MAIN model: A heuristic approach to understanding technology effects on credibility". In: M.J. METZGER and A.J. FLANAGIN (Eds.). *Digital media, youth, and credibility*. Cambridge, MA: The MIT Press, 73-100.
<https://doi.org/10.1162/dmal.9780262562324.073>

TANDOC Jr, E.C., YAO, L.J. and WU, S. (2020). "Man vs. Machine? The Impact of Algorithm Authorship on News Credibility". *Digital Journalism*, 8 (4), 548-562.
<https://doi.org/10.1080/21670811.2020.1762102>

TEJEDOR-CALVO, S., CERVI, L., PULIDO, C.M. and PÉREZ-TORNERO, J.M. (2021a). "Análisis de la integración de sistemas inteligentes de alertas y automatización de contenidos en cuatro cibermedios". *Estudios sobre el mensaje periodístico*, 27 (3), 973-983.
<https://doi.org/10.5209/esmp.77003>

TEJEDOR, S., CERVI, L., TUSA JUMBO, F. and PORTALÉS, M. (2021b). "La información de la pandemia de la Covid-19 en las portadas de los diarios. Estudio comparativo de Italia, Reino Unido, España, Francia, Portugal, Estados Unidos, Rusia y Alemania". *Revista Mexicana de Ciencias Políticas y Sociales*, 66 (242), 251-291.
<https://doi.org/10.22201/fcpys.2448492xe.2021.242.77439>

TÚÑEZ-LÓPEZ, M., TOURAL-BRAN, C. and VALDIVIEZO-ABAD, C. (2019). "Automatización, bots y algoritmos en la redacción de noticias. Impacto y calidad del periodismo artificial". *Revista Latina de Comunicación Social*, 74, 1411-1433.
<https://doi.org/10.4185/RLCS-2019-1391>

TÚÑEZ-LÓPEZ, J. M., FIEIRAS CEIDE, C. and VAZ-ÁLVAREZ, M. (2021). "Impact of Artificial Intelligence on Journalism: transformations in the company, products, contents and professional profile". *Communication & Society*, 34 (1), 177-193.
<https://doi.org/10.15581/003.34.1.177-193>

TURING, A. (1950). "Computing Machinery and Intelligence". *Mind*, LIX (236), 433-460.
<https://doi.org/10.1093/mind/LIX.236.433>

UFARTE-RUIZ, M.J. and MANFREDI SÁNCHEZ, J.L. (2019). "Algoritmos y bots aplicados al periodismo. El caso de Narrativa Inteligencia Artificial: estructura, producción y calidad informativa". *Doxa Comunicación*, 29, 213-233.
<https://doi.org/10.31921/doxacom.n29a11>

VAN-DER-KAA, H. and KRAHMER, E. (2014). "Journalist versus news consumer: The perceived credibility of machine written news". In: *Proceedings of the Computation + Journalism conference*, Columbia University, New York, 24–25, October. Retrieved from <https://research.tilburguniversity.edu/en/publications/journalist-versus-news-consumer-the-perceived-credibility-of-mach>

WADDELL, T.F. (2018). "A Robot Wrote This? How perceived machine authorship affects news credibility". *Digital Journalism*, 6 (2), 236-255.
<https://doi.org/10.1080/21670811.2017.1384319>

— (2019). "Can an Algorithm Reduce the Perceived Bias of News? Testing the Effect of Machine Attribution on News Readers' Evaluations of Bias, Anthropo-

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

WÖLKER, A. and POWELL, T.E. (2021). "Algorithms in the newsroom? News readers' perceived credibility and selection of automated journalism". *Journalism*, 22 (2), 86-103.
<https://doi.org/10.1177/1464884918757072>

ZHENG, Y., ZHONG, B. and YANG, F. (2018). "When algorithms meet journalism: The user perception to automated news in a cross-cultural context". *Computers in Human Behavior*, 86, 266-275.
<https://doi.org/10.1016/j.chb.2018.04.046>

WU, Y. (2019). "Is automated journalistic writing less biased? An experimental test of auto-written and human-written news stories". *Journalism Practice*, 14 (8), 1008-1028.
<https://doi.org/10.1080/17512786.2019.1682940>