VISUAL REVIEW | Vol. 18, No. 7, 2026 | ISSN 2695-9631
International Visual Culture Review / Revista Internacional de Cultura Visual
https://doi.org/10.62161/revvisual.v18.6474

ARTIFICIAL INTELLIGENCE AND INFORMATION VERIFICATION AGAINST DIGITAL DISINFORMATION: PERFORMANCE, TRUST AND ETHICS.
Technical performance and public perception in automated fact-checking processes

ANA LUCÍA DE VEGA MARTÍN [1]
analucia.devega@universidadviu.com
[1]Universidad Internacional de Valencia, España

KEYWORDS
Artificial intelligence
Information verification
Disinformation
Social media
Public trust
Transparency
Media literacy

ABSTRACT
The proliferation of digital disinformation has driven the use of artificial intelligence (AI) systems in information verification. This study examines the technical performance of these tools and public trust in their application to fact-checking. Using a mixed-methods design, it analyzes 150 verification reports integrating AI and a survey of 600 social media users from Spain and Latin America. The results show high effectiveness in automated verification alongside moderate levels of trust, conditioned by algorithmic transparency and bias concerns. The study concludes that hybrid AI–human models constitute the most robust strategy for strengthening information verification.

RECIBIDO: 05/ 08 / 2026
ACEPTADO: 20/ 10 / 2026

1. Introduction

In contemporary digital society, the mass circulation of information through social networks and digital platforms has deeply transformed the way citizens access, consume and evaluate informational content. While this scenario has widened opportunities for participation and access to knowledge, it has also fostered a sustained increase in disinformation, understood as a phenomenon that compromises the veracity of public discourse and conditions individual and collective decision-making (Wardle & Derakhshan, 2021). Recent studies on the role of artificial intelligence in this context highlight current challenges and approaches to combating disinformation, emphasising not only technical detection capacities but also difficulties related to bias, transparency and governance of automated systems (Saeidnia et al., 2025).

Disinformation has become a structural problem within the global media ecosystem, with direct effects on institutional trust and democratic quality. Recent episodes, such as the spread of false content during the COVID-19 pandemic, revealed the informational vulnerability of broad social sectors and the difficulty of distinguishing between verified information and manipulated content in digital environments (Mian & Khan, 2020). This context has reinforced the need to strengthen media and information literacy, understood as a critical competence for accessing, analysing and producing information in an ethical and responsible manner (De Vega Martín, 2023).

Within this framework, information verification acquires a relevance that goes beyond its technical dimension and becomes embedded in the sphere of symbolic and cultural mediation of truth in digital environments. The growing automation of fact-checking processes not only alters the speed and scale of verification, but also introduces new forms of informational authority, in which algorithms and automated systems actively participate in the construction of what is considered reliable, true or legitimate. From this perspective, disinformation does not constitute solely a problem of the circulation of false content, but rather a phenomenon that challenges contemporary modes of information production, validation and visualisation within the digital media ecosystem.

1.1. Artificial intelligence and information verification

In response to these challenges, artificial intelligence (AI) has assumed a central role in information verification processes. Advances in machine learning and natural language processing have enabled the development of systems capable of identifying discursive patterns, analysing large volumes of content and assessing the reliability of information in real time (Kreps et al., 2022). Their incorporation into the field of fact-checking has significantly improved the speed and efficiency of detecting false or manipulated content (Zhou et al., 2022; Kreps & Monin, 2024).

In recent years, various initiatives have applied AI-based solutions to limit the spread of disinformation. In the Spanish context, projects such as *IVERES*, promoted by RTVE for the early detection of disinformative content (Rigall, 2024), or *VerificAudio*, developed by PRISA Media in collaboration with the Google News Initiative and aimed at identifying audio deepfakes (PRISA Media, 2025). At the international level, major technology platforms have incorporated automated models that combine algorithmic verification with cross-checked databases (García, Liu & Zellers, 2023).

1.2. Ethical and social challenges of automated verification

Despite its technical potential, the use of AI in information verification poses significant ethical and social challenges. Several studies warn that these systems may reproduce biases derived from the data on which they are trained, affecting the neutrality and fairness of results (Bender et al., 2021). Likewise, algorithmic opacity and the limited explainability of automated processes contribute to generating public distrust and call into question the social legitimacy of these tools (García et al., 2023).

From a communicative and educational perspective, the incorporation of AI into information management cannot be understood solely as a technological advance. It also represents a formative challenge that requires promoting critical understanding of algorithms, data ethics and the role of human oversight. In this sense, media and information literacy emerges as a key axis for fostering responsible use of technology and mitigating the risks associated with automated information verification.

In this regard, the incorporation of automated verification systems raises questions that directly affect the relationship between technology, trust and information culture. The partial delegation of verification to algorithmic systems reconfigures traditional criteria of credibility and displaces, at least in part, human judgement as the exclusive instance of informational validation. This transformation is particularly significant in a media environment characterised by the centrality of the visual, the accelerated circulation of multimodal content and the increasing opacity of technological processes. Analysing automated verification therefore entails examining not only its operational effectiveness, but also its impact on contemporary forms of understanding, interpreting and critically evaluating information by citizens.

1.3. Purpose and objectives of the study

Against this background, the present study aims for analyse the role of artificial intelligence in information verification against digital disinformation, addressing both its technical performance and public perception and levels of trust in its application. Specifically, the following objectives are proposed:

1. To examine the performance of AI-based information verification systems through the analysis of platforms and projects that integrate automated fact-checking tools.
2. To analyse the perception and level of trust of social media users in AI-based verification systems, considering the ethical and social factors that condition their acceptance.

The relevance of this research lies in offering an empirical approach that articulates the technical effectiveness of AI with its social legitimacy, providing evidence of the gap between automated performance and public trust. From this perspective, the study contributes to the academic debate on information verification in digital environments and highlights the importance of hybrid models, based on the complementarity between automation and human oversight, as well as media literacy strategies aimed at strengthening citizens' information competence.

The study situates AI-based information verification as a process in which technical effectiveness and human supervision are articulated in the construction of informational credibility. The combination of performance analysis and public perception provides empirical evidence on the relationship between automation, trust and the legitimacy of verification processes in digital environments.

2. Design and method

The research adopts a mixed-methods approach, with a descriptive and explanatory, non-experimental and cross-sectional design, suitable for analysing complex phenomena that combine technical and social dimensions, such as information verification based on artificial intelligence (AI). This approach makes possible to integrate the analysis of the performance of automated systems with the study of public perception and trust, offering an articulated view between technological effectiveness and social legitimacy (Creswell & Creswell, 2018; Silver, 2024).

The formal object of the study is the use of AI systems in information verification processes against digital disinformation, addressed from a dual analytical perspective. On the one hand, the functioning and technical performance of automated verification tools used by specialised platforms are examined. On the other hand, the attitudes, perceptions and levels of trust of social media users regarding the application of these technologies in the detection of false or manipulated content are analysed.

Given the descriptive and explanatory nature of the study, no causal hypotheses are formulated. Instead, an objectives- and variables-based approach is adopted, allowing the exploration of relevant patterns, trends and relationships between AI performance and social acceptance. The scope of the study is situated within the Spanish-speaking digital ecosystem, with particular attention to Spain and Latin America, both in terms of the verification platforms analysed and the origin of the survey participants.

The research is organised into two complementary phases (Table 1): (i) a documentary and content analysis of verification reports produced by platforms that integrate AI, aimed at characterising their performance, methods and limitations; and (ii) a survey of social media users designed to measure perception, knowledge and trust in automated verification. The literature review is used to support

theoretical contextualisation and the definition of analytical categories, rather than as an independent method, given the empirical nature of the study.

Table 1. Methodological approach of the study phases

Study phase	Method	Objective	Techniques used
Platform analysis	Qualitative	Evaluate the implementation of AI in fact-checking	Documentary review and content analysis
User survey	Quantitative	Measure perception and trust in AI applied to fact-checking	Structured questionnaire with Likert scale

Source: Authors' own elaboration.
Note: the documentary review is used as a supporting technique for theoretical contextualisation, not as an independent research method.

2.1. Phase 1: analysis of verification platforms using artificial intelligence

The first phase focuses on the documentary and content analysis of 150 verification reports published by platforms that incorporate AI tools into their fact-checking processes. The selection of platforms was based on inclusion and exclusion criteria defined according to the specialised literature and international standards for digital verification (Wardle & Derakhshan, 2021), ensuring the presence of verifiable methodologies for AI use, a consolidated track record and an explicit commitment to transparency in verification processes (Table 2).

Table 2. Selection criteria for news verification platforms

Selection criterion	Justification
Documented use of AI in fact-checking	Only platforms with verified methodologies for AI use were selected
Relevance in the digital ecosystem	Platforms accredited by the International Fact-Checking Network (IFCN) and other specialised organisations
Track record and reliability	Platforms with at least three years of documented verification activity were included
Geographical coverage	Platforms active in Spain and Latin America were included to ensure representativeness in the Spanish-speaking context
Access to verified databases	Ensures that AI tools can compare information with reliable sources in real time
Transparency of AI models	Platforms that do not explicitly describe the decision criteria of their algorithms were exclude

Source: Authors' own elaboration.

The reports analysed, corresponding to a twelve-month period (November 2024–October 2025), constitute the unit of analysis for this phase. The analytical procedure was organised around a previously defined category system, structured into three main dimensions (Table 3): AI efficiency, methods employed, and challenges and limitations. These dimensions made it possible to assess aspects such as accuracy in detecting disinformation, response time of automated systems, the use of natural language processing and deep learning techniques, as well as the identification of algorithmic biases, false positives and human oversight mechanisms.

Table 3. Dimensions and indicators of analysis for Phase 1

Analytical dimension	Specific indicators
AI efficiency	Accuracy in detecting disinformation; response time; volume of content processed
Methods employed	Use of natural language processing; neural networks; deep learning; hybrid systems with human oversight
Challenges and limitations	Presence of algorithmic biases; level of methodological transparency; identification of false positives and false negatives; limits of explainability

Source: Authors' own elaboration.

This analysis provided a systematic overview of the degree of automation in verification processes and of the main technical strengths and limitations of AI-based systems, offering an empirical basis for the subsequent interpretation of the survey results.

2.2. Phase 2: survey on public perception of AI in information verification

The second phase adopts a quantitative approach and aims to analyse public perception and the level of citizen trust in the use of artificial intelligence applied to information verification. In line with the objectives of the study, this phase focuses on three main variables: perceived effectiveness, level of knowledge about AI applied to fact-checking, and degree of trust in automated verification systems. Contextual variables related to news consumption habits and personal verification practices are also incorporated, as they are considered relevant for interpreting attitudes towards automated verification.

Data collection for the survey was carried out independently from the documentary analysis of reports, as a differentiated phase aimed at examining public perception of AI-based information verification.

2.2.1. Sample, sampling and scope of the study

The sample consisted of 600 social media users, selected through non-probabilistic quota sampling, based on criteria of age, country of residence, educational level and prior experience with fact-checking platforms (Table 4). This type of sampling was considered appropriate given the absence of a universal sampling frame for social media users and its capacity to ensure internal diversity in sociodemographic profiles and levels of digital literacy.

Table 4. Sample selection criteria

Criterion	Description and justification
Age	Participants aged between 18 and 65 years were included, the age range with the highest news consumption in digital environments (Ipsos, 2023)
Country of residence	Participants from Spain, Mexico, Argentina and Colombia were selected, countries with active IFCN-accredited fact-checkers and relevant disinformation phenomena
Educational level	Individuals with secondary, undergraduate and postgraduate education were included to analyse differences in trust in AI according to academic level.
Experience with fact-checking	A distinction was made between users who have used information verification platforms and those who are unaware of their existence.

Source: Authors' own elaboration.

The territorial scope of the study included Spain, Mexico, Argentina and Colombia, selected for their relevance within the Spanish-speaking media ecosystem and the presence of active information verification initiatives. The age range of participants was between 18 and 65 years, a segment that concentrates higher levels of news consumption in digital environments..

2.2.2. Instrument and questionnaire structure

Data collection was carried out using a structured questionnaire administered remotely through a digital form. The instrument consisted of 20 items, organised into four thematic blocks, designed to operationalise the variables of the study.

Block 1, devoted to sociodemographic data, made it possible to characterise the participants' profile in terms of age, country of residence, educational level and linkage to the fields of communication, journalism or technology (Table 5).

Table 5. Block 1. Sociodemographic data

Question	Response type
What is your age?	Multiple choice: (18–24), (25–34), (35–44), (45–54), (55–65)
In which country do you currently reside?	Multiple choice: Spain, Mexico, Argentina, Colombia
What is your highest level of education?	Multiple choice: Secondary education, Undergraduate, Postgraduate
Do you work or study in the field of communication, journalism or technology?	Yes / No

Source: Authors' own elaboration.

Block 2 focused on news consumption habits and verification practices, including frequency of news consumption on social media, ways of access to information, prior experience with false news and strategies adopted when suspecting disinformation (Table 6).

Table 6. Block 2. News consumption habits

Question	Response type
How often do you consume news on social media?	Multiple choice: Several times a day; Once a day; Several times a week; Rarely; Never
Through which channels do you most frequently consume news?	Multiple choice: Social media; News websites; Television; Radio; Printed press
Have you ever detected false news on social media?	Yes / No
What do you do when you suspect that a piece of news may be false?	Multiple choice: Search for official sources; Consult fact-checkers; Share the news without verifying; Do nothing

Source: Authors' own elaboration.

Block 3 addressed knowledge and trust in AI applied to fact-checking through dichotomous questions and five-point Likert-scale items, aimed at assessing perceived effectiveness of AI, the level of trust generated by these tools and the evaluation of complementarity between automated systems and human fact-checkers (Table 7).

Table 7. Block 3. Knowledge and trust in AI applied to fact-checking

Question	Response type
Do you know or have you used any fact-checking platform?	Yes / No
Do you think AI can detect false news more effectively than a human?	Likert scale (1 = Strongly disagree, 5 = Strongly agree)
How much trust do you have in AI tools for verifying news?	Likert scale (1 = No trust, 5 = High trust)
Do you think human fact-checkers should complement AI systems in detecting fake news?	Yes / No

Source: Authors' own elaboration.

Finally, Block 4 analysed perceptions of the effectiveness and limitations of AI in information verification, including acceptance of automated mechanisms on digital platforms, identification of perceived limitations (such as lack of transparency, algorithmic bias or technical issues) and expectations regarding the possible replacement of human fact-checkers (Table 8).

Table 8. Perceived effectiveness of AI in news verification

Question	Response type
Do you think social media platforms should incorporate AI to automatically identify and remove false news?	Yes / No
What do you think are the main limitations of AI in news verification?	Multiple choice: Lack of transparency; Algorithmic bias; Lack of reliable data; Technical problems
Do you think AI will completely replace human fact-checkers in the future?	Yes / No

Source: Authors' own elaboration.

2.2.3. Application procedure and quality control

Participant recruitment was carried out through a combination of dissemination on social media, distribution of the questionnaire via academic university databases and collaboration with fact-checking organisations, which facilitated the circulation of the instrument through their dissemination channels. The application of the questionnaire took place over a three-month period (between September and November 2025), following a staggered dissemination strategy to maximise the diversity of participant profiles.

To ensure the quality of the instrument, the questionnaire underwent a validation process through expert judgement, in which five specialists in digital communication and artificial intelligence assessed the clarity, relevance and coherence of the items. In addition, a pilot test was conducted with 20 participants, allowing verification of question comprehension, response consistency and the average completion time, estimated at between 5 and 7 minutes.

2.2.4. Sampling and ethical considerations

It was observed that a proportion of participants had professional links to the fields of communication, journalism or technology. This aspect was considered relevant for the purposes of the study, given the active role of these profiles in the creation, circulation and verification of informational content, and it was taken into account in the interpretation of the results.

The study was conducted in accordance with ethical principles applicable to non-experimental, minimal-risk social research. Participation was voluntary and anonymous, with no collection of personally identifiable data. Before starting the questionnaire, participants received clear information about the objectives of the study, the exclusively academic use of the data and their right to withdraw from the survey at any time. Responses were analysed in aggregate form, ensuring confidentiality and ethical data handling.

3. Fieldwork and data analysis

The fieldwork was carried out in two empirical phases, in accordance with the methodological design of the study: the analysis of verification reports produced by platforms that integrate artificial intelligence (AI), and the administration of a structured survey to social media users.

3.1. Phase 1: collection and analysis of verification reports

The first phase consisted of the analysis of 150 verification reports published by fact-checking platforms that incorporate AI into their processes, corresponding to the period between November 2024 and October 2025. Each report constituted the unit of analysis.

It should be noted that the quantitative values of technical performance collected in this phase (such as accuracy, verification time and content volume) are based on the metrics and descriptions reported in the analysed platforms' own reports and public documentation. Consequently, they reflect the performance declared in those materials and do not correspond to an independent audit conducted by the research team.

The documents were collected and organised in a database for systematic processing. The analysis was conducted using content analysis, applying a matrix of categories previously defined and aligned with the objectives of the study. Coding made it possible to record indicators related to technical

efficiency (declared accuracy, response times and volume of processed content), methods employed (type of technology used and degree of human intervention), and limitations identified in the reports themselves. The procedure combined systematic counting of observable variables with an analytical description of the verification technologies and approaches employed by the platforms.

3.2. Phase 2: administration and processing of the survey

The second phase was conducted through an online questionnaire administered to 600 social media users from Spain, Mexico, Argentina and Colombia over a three-month period (September–November 2025). Participant recruitment was carried out through digital channels and academic networks, with the aim of maximising diversity in sociodemographic profiles and informational experiences.

Once fieldwork was completed, data underwent a preliminary cleaning process prior to analysis. Statistical treatment was carried out using descriptive statistics, employing frequencies and percentages for categorical variables, and means and standard deviations for items measured using Likert scales. This analysis made it possible to identify general patterns of perception, knowledge and trust in AI applied to information verification, following the structure of the blocks defined in the questionnaire.

3.3. Considerations applied during fieldwork

During the development of the fieldwork, the participation conditions and data processing criteria established in the methodological design were applied, ensuring voluntary participation, anonymity of responses and aggregated treatment of the collected information.

4. Results

4.1. Results of Phase 1: analysis of AI-based verification reports

The analysis of the 150 verification reports made it possible to identify consistent patterns in relation to the technical performance of artificial intelligence–based systems, the technologies employed in fact-checking processes, and the main operational limitations reported by the platforms themselves.

4.1.1. Accuracy, speed and processing capacity

The results show that automated verification systems achieve high levels of accuracy in detecting disinformative content. As shown in Table 9, average accuracy stood at 92.8%, with a range between 87% and 98%, indicating relevant differences between platforms, although without pronounced declines that would compromise overall system effectiveness.

In terms of temporal efficiency, the average verification time per item was 4.1 seconds, with minimum values of 2.5 seconds and maximum values of 6.8 seconds. This variability indicates that, although all systems operate within reduced timeframes, differences are associated with the type of technology employed and the degree of human intervention required in certain verification processes.

Regarding processing capacity, the analysed platforms handled an average volume of 11,200 verified items per day, with figures reaching up to 19,700 daily items in the most highly automated systems. These data highlight the scalability of AI-based systems in addressing large volumes of information in contexts characterised by high content circulation.

Table 9. Evaluation of the technical characteristics of the platforms

Indicator evaluated	Average value	Observed range
Accuracy in fake news detection	92,8%	87%-98%
Average verification time per news item	4,1 seconds	2,5 - 6,8 seconds
Volume of news analysed per day	11.200 items	6.500 - 19.700

Source: Authors' own elaboration.

The presentation of indicators related to accuracy, speed and processing volume provides a contextualised overview of the operational scope of the automated verification systems analysed and enables systematic comparison of the different technological approaches employed by the platforms.

4.1.2. Technologies employed and associated performance levels

Analysis of the reports made possible to identify an uneven distribution of technologies used in verification processes (Table 10). Natural language processing (NLP) was employed by 91% of the analysed platforms, constituting the most widespread base technology. These platforms recorded an average efficiency of 89.7%, slightly lower than that achieved by more complex approaches.

Table 10. Technologies used in AI-based disinformation detection

Method employed	Percentage of platforms using it	Average efficiency (%)
Natural Language Processing (NLP)	91%	89,7%
Deep learning with neural networks	76%	92,4%
Hybrid models (AI + human fact-checkers)	69%	95,1%
Image analysis and deepfake detection	42%	86,3%

Source: Authors' own elaboration.

The observed distribution of technologies reveals relevant differences in the degree of automation and average performance of verification systems, providing an empirical basis for analysing the complementarity between automated approaches and human oversight.

Deep learning models based on neural networks, present in 76% of the platforms, achieved higher average efficiency (92.4%), indicating improved performance in identifying complex patterns and non-linear relationships within analysed content.

Hybrid models, combining automated systems with human supervision, were present in 69% of the platforms and showed the highest average efficiency values (95.1%). This group also exhibited lower dispersion in results, suggesting more stable behaviour in verification processes.

Image analysis and deepfake detection systems, identified in 42% of platforms, showed an average efficiency of 86.3%, making them the lowest-performing approach in relative terms. However, these systems were primarily applied to specific audiovisual content, which limits their direct comparability with methods focused on textual content.

4.1.3. Need for human oversight and operational limitations

Analysis of the verification reports revealed recurrent references to the need for human validation, particularly in cases involving the correction of false positives and false negatives, as well as in content with a high contextual load. The reports also describe limitations related to the explainability of algorithmic processes and the traceability of verification decisions, both in NLP-based systems and in deep learning models. In the case of multimodal content analysis, technical constraints were identified in the detection of complex manipulations, especially in low-quality audiovisual material or content with subtle alterations.

These results highlight that, even in contexts of high automated performance, information verification continues to depend on processes of contextualisation and validation that cannot be resolved exclusively through algorithmic systems.

4.2. Results of Phase 2: public perception of AI in information verification

4.2.1. Sociodemographic characteristics of the sample

Table 11 presents the sociodemographic distribution of the sample. In terms of age, the most represented groups correspond to the 25–34 age range (28%) and 35–44 (25%), followed by the 18–24 group (21%). Older age groups show lower representation, with 16% in the 45–54 range and 10% in the 55–65 range.

Table 11. Sociodemographic data of the sample

Variable	Categories	Percentage
Age	18-24 years	21%
	25-34 years	28%
	35-44 years	25%
	45-54 years	16%
	55-65 years	10%
Country of residence	Spain	35%
	Mexico	30%
	Argentina	20%
	Colombia	15%
Educational level	Secondary education	35%
	Undergraduate	52%
	Postgraduate	13%
Link to communication, journalism or technology.	Yes	27%
	No	73%

Source: Authors' own elaboration.

Regarding country of residence, participation is concentrated in Spain (35%) and Mexico (30%), while Argentina (20%) and Colombia (15%) show lower percentages, consistent with the established recruitment criteria. With respect to educational level, more than half of the sample holds undergraduate degrees (52%) and 13% hold postgraduate qualifications. Finally, 27% of participants report professional or academic links to the fields of communication, journalism or technology.

4.2.2. News consumption habits on social media

Table 12 shows a high frequency of news consumption on social media. Fifty-eight per cent of participants report consulting news several times a day and 24% at least once a day, compared with 18% who do so less frequently. In relation to exposure to disinformation, 72% state that they identify false news on a daily basis or several times a week. However, 38% acknowledge having shared news without verification. Among these participants, the most frequent reason was that the content aligned with their beliefs (62%), followed by perceived source reliability (21%) and lack of time to verify (11%).

Table 12. News consumption habits

Variable	Categories	Percentage
Frequency of news consumption on social media	Several times a day	58%
	Once a day	24%
	Several times a week	12%
	Rarely	6%
Frequency of identifying false news	Every day	26%
	Several times a week	46%
	Occasionally	20%
	Never	8%
Has shared news without verification	Yes	38%
	No	62%
Reason for sharing unverified news	Aligned with beliefs	62%
	Appeared reliable	21%
	Lack of time to verify	11%
	Did not consider it relevant	6%

Source: Authors' own elaboration.

Mean values and dispersion for these variables are presented in Table 13. News consumption frequency shows a high mean (M = 3.34; SD = 0.87), while frequency of identifying false news is slightly lower (M = 2.90; SD = 0.89). The variable relating to sharing unverified news shows a low standard deviation (SD = 0.49), indicating a relatively homogeneous distribution of this behaviour across the sample.

Table 13. Means and standard deviations – Block 2

Variable	Mean	Standard deviation (SD)
Frequency of news consumption on social media (1–4)	3.34	.87
Frequency of identifying false news (1–4)	2.90	.89
Has shared news without verification (1 = No, 2 = Yes)	1.38	.49
Main reason for sharing unverified news (1–4)	1.61	.78

Source: Authors' own elaboration.

4.2.3. Knowledge and trust in AI applied to fact-checking

Table 14 presents the results related to levels of knowledge and trust in AI applied to information verification. Sixty-one per cent of respondents report not being aware of the use of AI in fact-checking processes, while 18% state that they have used AI-based tools to verify news. The remainder fall into intermediate levels of familiarity. In terms of trust, 39% report high levels, 35% moderate levels and 26% low levels. Among reasons for distrust, lack of transparency in processes is the most frequently cited (51%), followed by concerns about potential algorithmic bias (28%).

Table 14. Knowledge and trust in AI applied to fact-checking

Variable	Categories	Percentage
Level of knowledge of AI in fact-checking	Not aware of AI use	61%
	Has heard of it but does not know how it works	21%
	Aware of AI use but has not used it	15%
	Has used AI-based verification tools	18%
Level of trust in AI applied to fact-checking	High trust (4–5 Likert)	39%
	Moderate trust (3 Likert)	35%
	Low trust (1–2 Likert)	26%
Reasons for distrust	Lack of transparency	51%
	Potential algorithmic bias	28%
	Distrust of automated verification	21%

Source: Authors' own elaboration.

The mean values and dispersion of these variables are detailed in Table 15. The level of knowledge shows a low mean (M = 1.75; SD = 0.96), while trust in AI is situated at an intermediate level (M = 3.07; SD = 1.11), with a relatively high standard deviation indicating heterogeneity in responses.

Table 15. Means and standard deviations – Block 3

Variable	Mean	Standard deviation (SD)
Level of knowledge of AI in fact-checking (1–4)	1.75	.96
Level of trust in AI applied to fact-checking (1–5)	3.07	1.11
Reasons for distrust in AI (1 = No distrust, 2 = Distrust)	1.51	.50

Source: Authors' own elaboration.

4.2.4. Perceived effectiveness and limitations of AI in information verification

Table 16 summarises perceptions of the effectiveness of AI in comparison with human verification. Forty-three per cent of participants consider AI to be more effective, while 29% state that human fact-checkers are indispensable and 28% identify hybrid models as the preferred alternative. Regarding

perceived limitations, lack of transparency in decision-making is most frequently cited (47%), followed by potential algorithmic bias (32%) and difficulties in interpreting irony, sarcasm or cultural context (21%).

Table 16. Perceived effectiveness of AI in news verification

Variable	Categories	Percentage
Perception of AI effectiveness in fact-checking	More effective than humans	43%
	Human fact-checkers are indispensable	29%
	Hybrid model (AI + humans) is the best option	28%
Main perceived limitations of AI	Lack of transparency in decision-making	47%
	Potential algorithmic bias	32%
	Inability to interpret irony, sarcasm or cultural context	21%
Belief that AI will replace human fact-checkers in the future	Yes	18%
	No	82%

Source: Authors' own elaboration.

Table 17 presents the means and standard deviations associated with these perceptions. Perceived effectiveness of AI shows a mean close to the midpoint of the scale (M = 1.99; SD = 0.81), while perceived limitations show a moderate mean (M = 1.47; SD = 0.50). Expectations of complete replacement of human fact-checkers by AI show a low mean (M = 1.18; SD = 0.38), indicating a strong concentration of negative responses.

Table 17. Medias y DT Bloque 4

Variable	Mean	Standard deviation (SD)
Perceived effectiveness of AI in fact-checking (1–3)	1.99	.81
Perceived limitations of AI (1 = No limitation, 2 = Limitation)	1.47	.50
Expectation that AI will replace human fact-checkers (1 = No, 2 = Yes)	1.18	.38

Source: Authors' own elaboration.

The predominant valuation of hybrid models and the low expectation of complete replacement of human fact-checkers suggest that social acceptance of automated verification is articulated around criteria of complementarity rather than full automation.

The results of Phase 2 characterise public attitudes towards automated verification as ambivalent, combining recognition of its operational usefulness with reservations related to transparency, contextual interpretation and the role of human judgement.

5. Conclusions and Discussion

5.1. Conclusions

The aim of this study was to analyse the role of artificial intelligence in information verification processes against digital disinformation, integrating technical performance, public trust and the ethical challenges associated with its implementation. Based on the analysis of 150 verification reports and a survey of 600 social media users, several conclusions can be drawn that help delineate the actual scope of these technologies within the contemporary information ecosystem.

First, the results confirm that artificial intelligence has become a highly effective tool for information verification from an operational perspective. Its capacity to process large volumes of information rapidly and to detect patterns associated with disinformation makes it a key resource for addressing the scale and speed of digital information flows. However, this effectiveness is uneven depending on content type and the degree of contextualisation required, indicating that technical performance alone does not guarantee fully reliable verification in all scenarios.

Second, the study shows that public trust in automated verification remains at moderate levels and is conditioned by structural factors such as limited understanding of how these systems function, perceptions of algorithmic opacity and concerns about the reproduction of bias. This finding highlight

that social acceptance of artificial intelligence in the informational domain depends not only on accuracy, but also on how its processes are communicated, understood and perceived as legitimate.

Third, and centrally, the results indicate that ethical challenges constitute a decisive axis in the integration of artificial intelligence into information verification. The recurrent need for human oversight, limitations in the explainability of algorithmic decisions and difficulties in auditing verification processes point to a deeper ethical issue: the distribution of responsibility in systems that directly influence the construction of public truth. From this perspective, ethics does not appear as a normative add-on, but as a structural condition for the legitimacy of automated fact-checking.

A combined reading of both phases leads to the conclusion that hybrid models, which combine artificial intelligence with human supervision, represent the most robust option from technical, social and ethical standpoints. These models allow the efficiency of automation to be harnessed while incorporating mechanisms of control, contextualisation and accountability, which are particularly relevant in complex or multimodal content.

The study concludes that artificial intelligence constitutes a strategic resource for strengthening information verification against digital disinformation, but is only socially acceptable and ethically sustainable when articulated around principles of transparency, explainability, human oversight and shared responsibility. The findings suggest that the main challenge of automated verification does not lie in incremental improvements in algorithmic precision, but in the construction of verification models that are socially comprehensible, ethically responsible and culturally legitimate. The future of automated fact-checking does not lie in replacing human judgement, but in building collaborative ecosystems in which technology amplifies (rather than substitutes) the critical and ethical capacities of individuals.

5.2. Discussion

The incorporation of artificial intelligence into information verification processes has introduced significant transformations in the way content reliability is assessed in digital environments. Beyond its operational contribution to combating disinformation, these systems affect credibility criteria and trust-building dynamics, particularly within an ecosystem characterised by accelerated content circulation and the growing presence of multimodal information.

The results of this study confirm that artificial intelligence has become a structural component of contemporary information verification processes, especially due to its ability to operate at scale and within timeframes compatible with the logic of social media circulation. The high technical performance observed in the analysis of verification reports is consistent with recent research highlighting the effectiveness of automated systems in identifying discursive patterns, cross-checking information and prioritising potentially disinformative content (Guo et al., 2022; Zhou et al., 2022).

However, the findings also qualify these contributions by showing that such effectiveness is conditional and uneven. In line with studies on multimodal automated verification, the results indicate that AI-based systems encounter greater difficulties when content requires contextual interpretation, deep semantic analysis or the detection of complex visual manipulations (Akhtar et al., 2023; Usman et al., 2023). This limitation reinforces the idea that technical performance cannot be assessed in isolation, but must be considered in relation to content type and communicative context.

The comparison between Phase 1 and Phase 2 results reveals a significant gap between technical capacity and public trust, a phenomenon widely documented in recent studies on social interaction with algorithmic systems. Despite strong operational performance, the survey shows moderate trust levels, shaped by limited understanding of system functioning and perceptions of process opacity. This tension has been extensively discussed in the literature on responsible artificial intelligence, which emphasises that technical accuracy does not ensure social legitimacy without transparency and explainability (Bender et al., 2021; Floridi & Cowls, 2022).

From an ethical perspective, the results suggest that the main challenges of automated fact-checking extend beyond algorithmic bias to encompass broader issues of responsibility, governance and accountability. The recurrent need for human validation identified in the analysed reports points to a structural ethical problem: the difficulty of assigning clear responsibility when informational decisions are mediated by automated systems. This issue has been highlighted in recent normative frameworks,

which identify human oversight and explainability as fundamental principles for the ethical use of artificial intelligence in sensitive contexts (UNESCO, 2021; Floridi, 2023).

From this standpoint, the ethical dimension of automated verification is not limited to bias mitigation or algorithmic transparency, but affects the very criteria by which informational truth is determined in digital environments. The integration of automated systems into fact-checking introduces implicit decisions about content prioritisation, interpretation and reliability parameters, positioning ethics as a structural—rather than accessory—component of AI-based information verification.

Based on the findings, this study argues that AI-based information verification should not be understood merely as a technical evolution of fact-checking, but as a process that redistributes informational authority within the digital ecosystem. Automation introduces new intermediaries in information validation and partially displaces traditional credibility criteria, requiring a reconsideration of the relationship between technology, editorial responsibility and public trust beyond purely operational frameworks.

The preference for hybrid models expressed in Phase 2 reinforces these interpretations. Previous research has shown that users display greater acceptance of automated systems when they function as support for human judgement rather than as autonomous substitutes, particularly in tasks affecting informational credibility and public debate (Pennycook & Rand, 2022; Tandoc et al., 2023). The results of this study provide empirical evidence supporting the viability of hybrid approaches as technically, socially and ethically more robust solutions.

The findings also engage with recent research on the transformation of fact-checking in the digital ecosystem, which highlights the need to integrate automation within broader frameworks of information governance. Studies focusing on European and Latin American contexts emphasise that the legitimacy of automated verification depends on its articulation with principles of transparency, editorial responsibility and media literacy (Gutiérrez-Caneda & Vázquez-Herrero, 2024). The results of this work reinforce this perspective by showing that public trust is built not only on technical effectiveness, but also on ethical understanding and perception of these systems.

From an educational and communicative perspective, the low level of knowledge about how artificial intelligence functions detected in the sample confirms that the fight against disinformation cannot rely exclusively on technological solutions. Recent literature has demonstrated that interventions focused on media literacy, critical reflection and attention to accuracy reduce the spread of false content even in highly polarised environments (Pennycook & Rand, 2022; Reuters Institute, 2024). In this sense, the results reinforce the need to integrate AI within broader educational strategies aimed at strengthening citizens' media competence.

Overall, the findings show that artificial intelligence offers significant opportunities to reinforce information verification, but also poses ethical and social challenges that cannot be resolved solely through technical improvements. The consolidation of automated fact-checking as a legitimate practice depends on its integration into hybrid, transparent and socially comprehensible models capable of balancing efficiency, responsibility and public trust.

5.3. Scope of the study and directions for future research

The results of this study should be interpreted in light of its methodological framework. In Phase 1, the analysis focused on verification reports published by platforms that explicitly disclose the use of artificial intelligence in their processes, enabling the examination of documented and technically comparable materials. This methodological choice responds to the objective of analysing verifiable and transparent practices in automated fact-checking. The technical performance metrics analysed should be understood as documentary indicators of performance as reported by the platforms themselves, suitable for comparative characterisation of approaches and practices, but not equivalent to experimental measurements or independent external audits.

In Phase 2, the survey was designed using quota sampling to capture diverse profiles of social media users in Spanish-speaking contexts. This strategy, consistent with the descriptive nature of the study, made possible to explore public perception and trust in automated verification in real digital environments, without the aim of statistical generalisation to the entire population.

Based on the findings, future research could extend the analysis to other geographical and media contexts, as well as incorporate longitudinal designs to observe changes in trust and ethical perceptions

of AI-based fact-checking over time. Further research avenues include examining the impact of different strategies of algorithmic explainability, human oversight and accountability mechanisms on the social legitimacy of these systems.

The study highlights the need to continue exploring multimodal verification and the role of educational interventions and platform design in strengthening media literacy and the critical appropriation of automated verification tools.

References

Akhtar, M., Schlichtkrull, M., Guo, Z., Cocarascu, O., Simperl, E., & Vlachos, A. (2023). Multimodal automated fact-checking: A survey. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 5430–5448. https://doi.org/10.18653/v1/2023.findings-emnlp.361

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE.

Floridi, L. (2023). *Ethics, governance, and policies for artificial intelligence*. Springer. https://doi.org/10.1007/978-3-031-21448-6

Guo, Z., Schlichtkrull, M., & Vlachos, A. (2022). A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics, 10*, 178–206. https://doi.org/10.1162/tacl_a_00454

Gutiérrez-Caneda, B., & Vázquez-Herrero, J. (2024). Fact-checking y automatización en el ecosistema informativo digital: retos éticos y profesionales. *Profesional de la Información, 33*(1), e330104. https://doi.org/10.3145/epi.2024.ene.04

Mian, A., & Khan, S. (2020). Coronavirus: The spread of misinformation. *BMC Medicine, 18*, 89. https://doi.org/10.1186/s12916-020-01556-3

Pennycook, G., & Rand, D. G. (2022). Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature Communications, 13*, 2333. https://doi.org/10.1038/s41467-022-30073-5

Reuters Institute for the Study of Journalism. (2024). *Digital News Report 2024*. University of Oxford.

Saeidnia, H. R., Hosseini, E., Lund, B., Alipour Tehrani, M., Zaker, S., & Molaei, S. (2025). Artificial intelligence in the battle against disinformation and misinformation: A systematic review of challenges and approaches. *Knowledge and Information Systems, 67*(4). https://doi.org/10.1007/s10115-024-02337-7

UNESCO. (2021). *Recommendation on the ethics of artificial intelligence*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000380455

Vallayil, M., Nand, P., Yan, W. Q., & Allende-Cid, H. (2023). Explainability of automated fact verification systems: A comprehensive review. *Applied Sciences, 13*(23), 12608. https://doi.org/10.3390/app132312608

Wardle, C., & Derakhshan, H. (2021). *Information disorder: Toward an interdisciplinary framework for research and policymaking*. Council of Europe.

Zhou, X., Wu, J., & Zafarani, R. (2022). SAFE: Similarity-aware fact checking. *Proceedings of the Web Conference 2022*, 1896–1906. https://doi.org/10.1145/3485447.3512244