The AI Trust Paradox in Journalism

Navigating Perception, Biometrics, and Economic Viability

Barbara Brandstetter
Information Management
HNU University of Applied Sciences
Neu-Ulm, Germany
email: barbara.brandstetter@hnu.de

Anja Zenk
Information Management
HNU University of Applied Sciences
Neu-Ulm, Germany
email: anja.zenk@hnu.de

Abstract - **This paper investigates the "AI Trust Paradox" in journalism, in which the integration of artificial intelligence offers unprecedented efficiency yet simultaneously triggers public scepticism. Using a multi-method experimental approach that combines biometric facial-expression data and self-report surveys, this study examines how audiences perceive and evaluate AI-generated journalistic content. The findings reveal a systematic divergence between subconscious and conscious responses: while participants consciously rate AI-generated content lower in credibility, entertainment value, and willingness to pay than human-authored content, they subconsciously express something different. Misclassification analyses demonstrate that these differences are driven by perceived authorship attribution rather than objective content quality. The paper concludes that media organisations must carefully navigate the tension between the normative demand for transparency and its adverse effects on perceived credibility and economic viability.**

Keywords - AI in journalism; algorithm aversion; elaboration likelihood model; biometrics; willingness to pay.

I. INTRODUCTION

The integration of Artificial Intelligence (AI) is fundamentally transforming publishing houses, reshaping how news is produced, distributed, and consumed [1][2]. Media organisations increasingly rely on automated systems for routine reporting, freeing journalists to focus on complex, contextual, and investigative tasks. Prominent broadcasters such as Bayerischer Rundfunk and the German Press Agency (dpa) have already established formal AI guidelines to ensure editorial transparency [3][4]. This rapid adoption creates a profound tension referred to here as the AI Trust Paradox: For media organisations, AI represents a tool for efficiency [5]; for audiences, however, any perceived AI involvement tends to trigger generalised scepticism [6][7].

Empirical research consistently shows that consumers cannot reliably distinguish human-authored from AI-generated content based on perceptual cues alone [8][9]. Nevertheless, the mere disclosure of AI involvement substantially reduces perceived credibility, authenticity, and willingness to pay [11][12]. However, recent systematic reviews highlight that this "AI penalty" is not universal; rather, audience trust is highly contingent on how AI provenance and transparency cues are designed and on whether human oversight is communicated [10].

This study employs a multi-method experimental design, combining biometric facial expression analysis with post-stimulus questionnaires, to examine the emotional and cognitive mechanisms underlying audience responses to AI-generated journalism. Three research questions guide the inquiry:

RQ1: How do audiences perceive and evaluate AI-generated content compared to human-created journalistic content with respect to quality, credibility, and authenticity?

RQ2: To what extent can audiences distinguish between AI-generated and human-authored content, and how does prior knowledge about AI authorship influence trust?

RQ3: How does AI authorship - and its disclosure - affect audience engagement, emotional responses, and willingness to pay (WTP)?

The paper is structured as follows: Section II reviews the related literature on AI in journalism, algorithm aversion, and the disclosure penalty. Section III describes the experimental methodology, including stimuli and data collection procedures. Section IV shows the results organised by research questions, covering both self-reported evaluations and biometric findings. Section V reports the multi-method triangulation, including the attribution effect and the isolated disclosure penalty. Section VI summarises the answers to the three research questions. Section VII discusses the theoretical contributions and practical implications, Section VIII provides the Conclusion, and Section IX discusses the limitations of the study.

II. RELATED LITERATURE

The integration of AI into journalism must be examined not only in terms of workflows but also in terms of its impact on public trust, perceptions of authenticity and the WTP.

A. AI Integration and Journalistic Roles

Analysed through the lens of Human-Machine Communication (HMC), machines are increasingly assuming communicative roles traditionally reserved for human journalists [12][13][14]. AI systems excel at processing structured data, generating standardised reports, and producing audio content - capabilities that are already being deployed by publishers globally [7]. The implementation of AI in core journalistic workflows necessitates a fundamental

reassessment of the relationship between journalists and automated systems [15]. While AI tools are widely adopted for news curation, content production, and audience interaction, journalists tend to resist growing technological dependency to protect their professional autonomy and editorial identity [16]. Although AI enhances operational efficiency, human journalists are increasingly positioned as providers of verification, contextualization, and empathetic storytelling - capacities that current AI systems cannot replicate [17][18].

B. Public Trust, Algorithm Aversion, and Authenticity

Public trust in journalism is grounded in perceived credibility, objectivity, and accuracy [6]. According to Mayer, Davis, and Schoorman's integrative model of organisational trust [19], trustworthiness rests on three pillars: ability (domain-specific competence), benevolence (the belief that the trustee acts in the trustor's interest), and integrity (adherence to shared principles) [19].

AI complicates this framework by triggering profound psychological resistance, most notably algorithm aversion. (In this paper, algorithm aversion is used in a broader meaning as a general scepticism towards the use of AI). When applied to automated systems, audiences perceive AI as an opaque "black box" incapable of genuine benevolence or moral integrity, given its absence of human empathy and conscious will [6][20]. Humans hesitate to trust machines because interpersonal trust traditionally requires attributing a conscious "will" to the trustee [20]; machines fail this attribution by design.

This trust deficit can be elucidated through the Elaboration Likelihood Model (ELM) of persuasion [22]. When audiences lack algorithmic literacy, explicit AI disclosure operates as a negative peripheral cue, triggering immediate scepticism and rejection [23]. Even in contexts of central-route processing, pre-existing concerns about algorithmic bias produce biased elaboration: audiences selectively amplify information that confirms their distrust [6].

C. The Distinguishability Problem and the Disclosure Penalty

Research consistently demonstrates that audiences cannot reliably differentiate between human-authored and AI-generated content on perceptual grounds alone [8]. This finding raises a fundamental paradox: while AI-generated content is indistinguishable in quality, explicitly labelled AI content receives markedly lower trust ratings. It is precisely this normatively required transparency that triggers the “disclosure penalty” which poses a direct threat to the economic viability of AI-augmented journalism [24].

However, a recent systematic review by Licenji and Hoxha emphasises the crucial distinction between AI provenance (the actual use of AI in production) and AI disclosure cues (the transparency labels provided to audiences). Their analysis reveals that negative effects on credibility and trust are not inevitable, but depend heavily on the specific design of the disclosure [10]. Disclosures that imply full automation without signalling editorial accountability or human oversight are particularly likely to trigger audience scepticism [10]. Consumer research suggests that human labour is intrinsically valued, and that the automation of creative tasks is perceived as diminishing authenticity [25]. When audiences learn of AI involvement, they infer a lack of human passion and commitment, leading to attitudinal and behavioural backlash that reduces both credibility assessments and purchase intentions [12].

III. METHODOLOGY

The following section outlines the methodological framework employed to investigate how audiences perceive and respond to AI-generated versus human-produced journalistic content across different media formats. By combining biometric measurements with self-reported evaluations in a controlled experimental setting, the study aims to capture both implicit emotional reactions and explicit assessments, thereby providing a multidimensional perspective on trust, credibility, and WTP in AI-augmented journalism.

A. Research Design and Sample

This study employs a quantitative, multi-method experimental design conducted in the usability laboratory at our university, using the software iMotions. The experimental approach allows for controlled exposure to stimuli and the simultaneous capture of both subconscious facial-expressive responses such as joy and engagement and conscious self-reported evaluations. The sample comprises 28 purposively selected participants.

It should be noted that the biometric and survey data originate from partially overlapping but non-identical subsamples. All 28 participants completed the post-stimulus questionnaires, yielding $n = 347$ responses across 13 stimuli (364 possible responses; response rate: 95.3%). Biometric data from iMotions facial expression analysis (AFFDEX) were successfully recorded for 20 of these 28 participants, producing $n = 208$ stimulus-level observations. The discrepancy is attributable to technical recording failures in the usability laboratory (e.g., insufficient facial landmark detection due to glasses, head angle, or lighting conditions) and does not reflect systematic exclusion. All 28 participants are included in the survey-based analyses; the biometric analyses draw on the subset of 20 participants for whom valid continuous recordings were available across all stimuli. Findings from the two data streams should therefore be interpreted with this difference in mind: while the survey results reflect the full sample of 28, the biometric results reflect a subsample of 20, which further underscores the exploratory character of the biometric findings.

B. Stimuli

To reflect the multimodal character of contemporary journalism, participants were exposed to 13 stimuli across three content formats: videos, audio podcasts, and written news articles. Stimuli were drawn from established German media outlets and assigned to one of three conditions: (1) Human - journalist-authored content without AI involvement; (2) AI Disclosed - AI-generated content

carrying an explicit disclosure label, replicating current editorial practice at outlets such as Bayerischer Rundfunk [3] and dpa [4]; and (3) AI Undisclosed - AI-generated content presented without any disclosure, simulating a non-transparent production scenario. Table I provides an overview of all stimuli.

TABLE I. OVERVIEW OF EXPERIMENTAL STIMULI

Format	**Stimuli Overview**		
	Topic / Genre	***Source***	***Condition***
Video	Financial	Business Insider	AI Disclosed
Video	Financial	Finanztip	Human
Video	Regional	WLZ Nachrichten	AI Undisclosed
Video	Regional	Studio 47	AI Disclosed
Video	Regional	Studio 47	AI Disclosed
Video	Regional	SWR	Human
Audio	Politics	Podcast Ronzheimer	Human
Audio	Politics	Podcast Ronzheimer (EN)	AI Disclosed
Audio	Politics	Upday	Human
Audio	Politics	Upday	AI Undisclosed
Text	Misc. News	Kölner Express	AI Disclosed
Text	Misc. News	Augsburger Allgemeine	Human
Text	Misc. News	Kölner Express	AI Undisclosed

Conditions: Human = journalist-authored; AI Disclosed = labeled AI-generated; AI Undisclosed = unlabeled AI-generated.

Data collection proceeded in two sequential phases. In the first phase, participants were exposed to each stimulus individually while continuous biometric data were recorded. Facial expression analysis was conducted using iMotions software integrated with the Affectiva AFFDEX engine [26], [27] which captured seven basic emotions (joy, sadness, anger, disgust, fear, surprise, contempt), valence, and engagement at the frame level. Emotion metrics are operationalised as the percentage of recording time during which a given emotion exceeded a detection threshold of 25 probability units, as determined using the R-based FEA Thresholding pipeline [26].

In the second phase, participants completed a short post-stimulus questionnaire immediately following each exposure. Six Likert-scaled items (1 = strongly disagree, 5 = strongly agree) assessed perceived informativeness, credibility, professionalism, entertainment value, enjoyment of consumption, and situational WTP. Participants additionally indicated whether they believed the content had been produced with AI assistance (yes/no) and, if so, provided an open-text justification.

C. Analytical Strategy

Biometric time-series data were aggregated to one observation per participant per stimulus, yielding n = 208 stimulus-level observations. With 20 participants and 13 stimuli, this results in 260 observations. The discrepancy of 52 missing observations is due to a tracking error rate. Survey data were pooled across all stimuli, producing n = 347 post-stimulus questionnaire responses from 28 participants across 13 stimuli (364 possible responses; response rate: 95.3%).

Correct AI detection rates are calculated as the proportion of participants providing the objectively accurate response at the stimulus level. All findings are classified as exploratory and interpreted as hypothesis-generating rather than confirmatory.

IV. RESULTS

The analysis integrates two complementary data streams: biometric measurements from iMotions AFFDEX facial expression analysis (n = 208 stimulus observations) and self-reported post-stimulus survey data (n = 347 responses). The following sections present findings organised by research questions.

A. RQ1: Perception and Evaluation of Content Quality

Self-reported evaluations revealed consistent differences across key quality dimensions (Table II). Human-generated content received the highest mean ratings on all six items. Credibility declined from M = 3.80 (Human) to M = 3.28 (AI Disclosed) and M = 3.41 (AI Undisclosed). Entertainment value showed the steepest decline, from M = 3.09 (Human) to M = 2.61 (AI Disclosed) and M = 1.99 (AI Undisclosed). Willingness to pay was higher for human content (M = 1.94) than for AI Disclosed (M = 1.65) or AI Undisclosed (M = 1.63).

TABLE II. SELF-REPORTED EVALUATIONS BY CONDITION (MEAN LIKERT 1-5)

Self-reported evaluations by condition			
Dimension	***Human***	***AI Disclosed***	***AI Undisclosed***
Credibility	3.80	3.28	3.41
Informativeness	3.72	3.44	3.59
Professionalism	3.49	3.07	2.71
Entertainment value	3.09	2.61	1.99
Enjoyment	2.95	2.51	2.22
Willingness to pay	1.94	1.65	1.63

n = 347 responses. Scale: 1 = strongly disagree, 5 = strongly agree.

In the video format, credibility declined from M = 4.18 (Human) to M = 2.89 (AI Undisclosed) - the most pronounced gap across all format–condition combinations - and WTP dropped from M = 2.20 to M = 1.18, representing a near halving. Qualitative explanations for these declines concentrated on visual AI cues: participants cited unnatural facial movements, repetitive head gestures, and the obvious avatar appearance of AI-generated video moderators. The text format displayed an unexpected reversal: AI Undisclosed text received higher credibility ratings (M =

3.58) than Human text (M = 2.48). This pattern is attributable to the very low AI detection rate for written content (46%) in this sample, which prevented label-triggered aversion from activating. These format-specific results suggest that perceptually salient AI artefacts in video substantially amplify the disclosure penalty, while written AI cues are insufficiently distinctive to elicit the same rejection response.

B. RQ2: Distinguishability and Transparency Demand

Overall, correct AI detection rates were 80% for AI Disclosed and 73% for AI Undisclosed content. These aggregate figures, however, conceal extreme variability at the stimulus level (Table III). The WLZ Nachrichten avatar moderator was correctly identified by 100% of participants, reflecting the perceptually obvious nature of current AI-generated video avatars. By contrast, an article about the Duchess of Sussex, Meghan, published by Kölner Express– (AI Disclosed text) achieved a correct classification rate of only 31%, despite carrying an explicit AI label - suggesting that label comprehension or salience failures are consequential in text formats. Most theoretically significant was the Augsburger Allgemeine misclassification: 84% of participants incorrectly identified this Human-authored text about Meghan as AI-generated. This high false-positive rate confirms the Distinguishability Problem [7][8]: AI-typical stylistic cues (generic framing, absence of editorial voice, schematic structure) can trigger false AI attribution in human-authored content, demonstrating that the issue operates in both directions.

TABLE III. CORRECT AI DETECTION RATES BY STIMULUS

Correct AI Detection Rates By Stimulus			
Stimulus	***Condition***	***Format***	***Correct %***
SWR (Music clubs)	Human	Video	93%
Ronzheimer (DE)	Human	Audio	88%
Upday (Harbour)	Human	Audio	70%
Finanztip	Human	Video	52%
Augsburger Allg.	Human	Text	16%
Ronzheimer (ENG)	AI Disclosed	Audio	85%
Studio 47 (Transformation)	AI Disclosed	Video	96%
Studio 47 (Feuer)	AI Disclosed	Video	96%
Kölner Express (Meghan)	AI Disclosed	Text	31%
Business Insider	AI Disclosed	Video	92%
WLZ Nachrichten	AI Undisclosed	Video	100%
Upday (Minimum wage)	AI Undisclosed	Audio	72%
Kölner Express (Drugs)	AI Undisclosed	Text	46%

Detection rate per stimulus

Demand for disclosure transparency was uniformly high and condition-independent: the perceived importance of knowing specifically how AI was used was rated M ≈ 4.5 on a 5-point scale across all three conditions. This finding indicates that the normative expectation of disclosure does not depend on whether AI was actually used - it is a stable audience expectation with direct implications for editorial policy.

C. RQ3: Emotional Engagement and Willingness to Pay

The biometric data collected reveal a pattern that systematically diverges from self-reported evaluations. The following analysis is exploratory in character. Descriptively, AI Undisclosed content produced the highest mean joy time percentage (5.66% of frames above threshold) and engagement time percentage (31.47%), compared to Human (joy: 3.84%, engagement: 27.41%) and AI Disclosed (joy: 3.99%, engagement: 27.98%). In the video format, this pattern was especially pronounced: AI Undisclosed stimuli elicited a joy time percentage of 13.8% and an engagement time percentage of 38.3%, substantially exceeding Human video benchmarks (joy: 3.5%, engagement: 26.9%).

A distinction between detection rate and conditional intensity proved analytically important. The median percentage time was 0% across most observations, indicating that detectable emotional expression was absent in the majority of stimulus exposures. Joy was detected in 52.5% of Human observations but only 43.8% of AI Undisclosed observations; fear detection rates were comparable across conditions (Human: 23.8%, AI Undisclosed: 22.9%). However, when restricting analysis to observations where a given emotion was detected at all, AI Undisclosed content showed markedly higher conditional intensity (joy: 12.9% vs. 7.3% for Human; fear: 6.8% vs. 2.8%). AI Undisclosed content thus elicited emotion less frequently, but more intensely when it did occur.

V. MULTI-METHOD TRIANGULATION

The following results highlight a central tension in audience responses to AI-generated journalism, revealing a systematic divergence between implicit emotional reactions and explicit evaluative judgments.

A. Convergences and Divergences

The central contribution of this study's multi-methods design is the systematic comparison of biometric and self-reported data, which reveal divergences invisible to either method alone (Table IV). The divergence is most pronounced for AI Undisclosed content, which simultaneously produced the highest subconscious emotional activation - the highest joy and the highest engagement - and the lowest conscious ratings for credibility, entertainment, and willingness to pay. This bidirectional divergence constitutes direct empirical evidence for the AI Trust Paradox as a psychophysiological phenomenon: audiences are affectively more engaged by content whose AI origin they are unaware of, yet assign it substantially lower

evaluative ratings once AI authorship is attributed. The pattern is consistent with the ELM framework [22]: without AI attribution, audiences process content via the central route and respond to its intrinsic affective qualities; once AI authorship is known, the label activates a negative peripheral cue that systematically overrides content-based evaluation [22].

TABLE IV. MULTI-METHODS TRIANGULATION CONDITION

Multi-Method Triangulation Condition			
SURVEY (Ø Likert 1–5)			
Measure	***Human***	***AI Disclosed***	***AI Undisclosed***
Credibility	3.80	3.28	3.41
Willingness to pay	1.94	1.65	1.63
BIOMETRICS (Ø frames with detected emotion, AFFDEX threshold 25)			
Joy (% of frames)	3.84%	3.99%	5.66%
Engagement (% of frames)	27.41%	27.98%	31.47%

B. The Attribution Effect: Misclassification Evidence

To isolate the causal role of attribution from that of content quality, within-condition misclassifications were analysed. When participants incorrectly identified human-generated content as AI-produced (n = 47), credibility dropped from M = 4.20 to M = 3.06, enjoyment from M = 3.40 to M = 2.15, and WTP from M = 2.15 to M = 1.55. The same content received dramatically lower evaluations solely based on its perceived - not actual - authorship. This indicates that attribution, not content, drives the devaluation of AI-generated journalism.

The symmetric case is equally revealing. Participants who incorrectly believed AI Undisclosed content to be human-authored (n = 21) rated it at M = 3.95 for credibility and M = 2.43 for WTP - the highest willingness to pay of any subgroup in the study, exceeding even the correctly attributed Human content group (M = 2.15). This effect confirms that AI content quality is fully capable of eliciting evaluative responses comparable to those of human journalism, provided that AI attribution is not activated.

VI. SUMMARY OF FINDINGS

Taken together, the findings demonstrate that audience evaluations of journalistic quality, credibility, and economic value are driven less by the content itself than by its perceived authorship

A. Quality, Credibility, and Authenticity

The findings show that AI-generated or assumed AI-generated content is evaluated less favourably than human-generated journalism across all core quality dimensions, with human-authored content receiving the highest ratings on credibility, professionalism, entertainment, and willingness to pay. Most importantly, misclassification analyses indicate that this depreciation is attributional rather than content-based: identical material is rated substantially higher when perceived as human-authored and lower when perceived as AI-generated.

B. Distinguishability and Knowledge Effects

Participants correctly identified AI authorship in 80% of disclosed and 73% of undisclosed exposures, indicating non-trivial detection capability with substantial stimulus-level variability. High false-positive rates for human-authored content, such as the Augsburger Allgemeine case, illustrate the bidirectional nature of the Distinguishability Problem [7][8]: audiences apply AI-attribution schemas based on unreliable stylistic cues.

C. Emotional Engagement and Economic Implications

The joint analysis of biometric and survey data reveals a paradox that would have been invisible to either method in isolation. At the subconscious level, AI Undisclosed content generated the highest affective activation, whereas at the conscious level it received the lowest ratings for entertainment, enjoyment, and willingness to pay. This divergence shows that AI-generated content does not inherently suppress emotional engagement; rather, evaluations vary with perceived authorship.

VII. DISCUSSION

The present findings extend and refine the existing literature on the AI disclosure penalty [24] and algorithm aversion [21] by providing integrated biometric–survey evidence of the AI Trust Paradox at the level of individual stimulus exposures.

The severe devaluation observed in our study when generic AI labels were present confirms that transparency in journalism must be treated as a complex design and communication problem, rather than a simple binary choice [10]. Three theoretical contributions are particularly notable.

First, the misclassification evidence - showing that identical content receives dramatically different ratings depending on perceived authorship. Erroneous AI attributions of human-authored content systematically reduce perceived value, whereas erroneous human attributions of AI-generated content correspondingly inflate it.

Second, the divergence between biometric and survey data provides psychophysiological evidence of the AI Trust Paradox. The finding that subconscious affective engagement is highest for AI Undisclosed content, while conscious evaluation is lowest, suggests that the peripheral cue function of AI labels operates as a post hoc override of affective processing - not as a determinant of initial engagement.

Third, the format-specific moderation of these effects has practical implications for editorial strategy. The disclosure penalty is most severe in video formats, where AI artefacts are visually salient, and weakest or reversed in text formats, where AI stylistic cues are less reliably identified. This asymmetry suggests that disclosure policies should be calibrated to format-specific audience expectations and detection capabilities.

VIII. CONCLUSION

This study provides exploratory evidence for the AI Trust Paradox in journalism. Three principal conclusions emerge:

(1) Attribution of AI authorship systematically depresses conscious audience evaluations of journalistic content, rather than reflecting intrinsic quality differences.

(2) Audiences show limited perceptual ability to distinguish AI-generated from human-authored content, while maintaining a strong normative demand for transparent disclosure.

(3) Biometric data reveal a divergence between subconscious emotional activation and conscious evaluation, indicating that AI labels can reshape audience responses post hoc.

These findings present media organisations with a strategic challenge that cannot be resolved by transparency or opacity alone. Addressing this challenge requires investment in audience education about AI's role in journalism, the development of disclosure formats that convey human editorial oversight rather than wholesale AI replacement, and a clear strategic differentiation between AI-assisted operational tasks and premium, human-driven editorial content. The distinctive value of human journalism - contextual judgment, empathetic storytelling, and democratic accountability - must remain visible and credibly communicated to audiences navigating an increasingly AI-augmented media environment.

IX. LIMITATIONS

This study is subject to several limitations that define the scope of its conclusions and point to productive directions for future research. The purposively selected sample of N = 28 limits statistical power and precludes population-level generalisation; the findings should be treated as exploratory, pending replication in larger - randomly sampled - populations. The focus on the German media market constrains cross-cultural generalisability, given documented international variation in AI acceptance and media trust. Self-reported survey data are susceptible to social desirability effects and introspective inaccuracies, while the controlled laboratory setting may not fully represent naturalistic media consumption conditions. Finally, the rapid evolution of generative AI technology means that audience responses and detection capabilities will change as AI-generated content becomes more prevalent and technically sophisticated, necessitating longitudinal monitoring of the phenomena documented here.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support of the usability laboratory technician of HNU, Stefan Rogg, and the participants who contributed their time to this research.

REFERENCES

[1] A. L. Guzman and S. C. Lewis, "What Generative AI Means for the Media Industries," Emerging Media, vol. 2, no. 3, pp. 347-355, 2024.

[2] L. Amigo and C. Porlezza, "Journalism Will Always Need Journalists: The Perceived Impact of AI on Journalism Authority in Switzerland," Journalism Practice, vol. 19, no. 10, pp. 2266–2284, 2025.

[3] K. Brunner, R. Ciesielski, P. Gawlik, U. Köppen, S. Kühne, J. Pfeiffer and C. Schneider, "Unsere KI-Richtlinien im Bayerischen Rundfunk,"(„Our AI Guidelines at Bayerischer Rundfunk“), Bayerischer Rundfunk, 2024.

[4] S. Raabe, "Offen, verantwortungsvoll und transparent – Die Guidelines der dpa für Künstliche Intelligenz," („Open, Responsible, and Transparent – The dpa’s Guidelines for Artificial Intelligence“), dpa, 2023.

[5] C. Beckett, "New Powers, New Responsibilities: A Global Survey of Journalism and Artificial Intelligence," London School of Economics, 2019.

[6] V. L. Sinclair, "The Influence of AI-Generated News on Public Trust in Journalism: Evidence from the UK," Journal of Research in Social Science and Humanities, vol. 4, no. 2, 2025.

[7] N. Newman, A. R. Arguedas, C. T. Robertson, R. K. Nielsen, and R. Fletcher, "Reuters Institute Digital News Report 2025," Reuters Institute for the Study of Journalism, 2025.

[8] Die Medienanstalten: "Transparenz-Check – Wahrnehmung von KI-Journalismus" (“Transparency Check – Perceptions of AI Journalism”), 2024.

[9] J. M. Chein., S. A. Martinez and A. R. Baron, „Human intelligence can safeguard against artificial intelligence: Individual differences in the discernment of human from AI texts.” Scientific Reports, 14. https://doi.org/10.1038/s41598-024-76218-y, 2024.

[10] L. Licenji and J. Hoxha, "When news is 'written by artificial intelligence': a systematic review of provenance and disclosure cues in journalism and their effects on credibility and trust," Frontiers in Artificial Intelligence, vol. 9, art. no. 1815243, 2026.

[11] J. D. Brüns and M. Meißner, "Do You Create Your Content Yourself? Using Generative Artificial Intelligence for Social Media Content Creation Diminishes Perceived Brand Authenticity," Journal of Retailing and Consumer Services, vol. 79, p. 103790, 2024.

[12] BSI Marketing Intelligence Lab. “The Impact of Artificial Intelligence on the Willingness to Pay for Digital Newspaper Subscriptions - An Empirical Study of the German News Market”. URL: https://www.bsi.ag/cases/145-case-studie-auswirkungen-des-einsatzes-kuenstlicher-intelligenz-auf-die-zahlungsbereitschaft-von-online-medien.html, 2025

[12] S. C. Lewis, A. L. Guzman, and T. R. Schmidt, "Automation, Journalism, and Human-Machine Communication: Rethinking Roles and Relationships of Humans and Machines in News," Digital Journalism, vol. 7, no. 4, pp. 409–427, 2019.

[13] A. K. Schapals and C. Porlezza, "Assistance or Resistance? Evaluating the Intersection of Automated Journalism and Journalistic Role Conceptions," Media and Communication, vol. 8, no. 3, pp. 16–26, 2020.

[14] H. Cools, B. van Gorp, and M. Opgenhaffen, "When Algorithms Recommend What's New(s): New Dynamics of Decision-Making and Autonomy in Newsgathering," Media and Communication, vol. 9, no. 4, pp. 198–207, 2021.

[15] M. Grimme and C. Zabel, "AI in the Newsroom: A Collective Case Study about Newsworker-AI Collaboration in the German Newspaper Industry," Journal of Media Business Studies, vol. 22, no. 2, pp. 118–142, 2025.

[16] P. Matich, T. J. Thomson, and R. J. Thomas, "Old Threats, New Name? Generative AI and Visual Journalism," Journalism Practice, vol. 19, pp. 1–22, 2025.

[17] L. A. Møller, A. van Dalen, and M. Skovsgaard., "A Little of that Human Touch: How Regular Journalists Redefine Their Expertise in the Face of Artificial Intelligence," Journalism Studies, vol. 26, no. 1, pp. 1–19, 2025. doi: 10.1080/1461670X.2024.2412212.

[18] C. Londoño-Proaño and J. Buele, "Can artificial intelligence replace journalists? A theoretical approach," Frontiers in Communication, vol. 10, 2025. doi: 10.3389/fcomm.2025.1537146.

[19] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An Integrative Model of Organisational Trust," Academy of Management Review, vol. 20, no. 3, pp. 709–734, 1995.

[20] J. De Freitas, S. Agarwal, B. Schmitt, and N. Haslam, "Psychological Factors Underlying Attitudes toward AI Tools," Nature Human Behaviour, vol. 7, pp. 1845–1854, 2023.

[21] B. J. Dietvorst, J. P. Simmons and C. Massey, “Algorithm aversion: People erroneously avoid algorithms after seeing them err,” Journal of Experimental Psychology: General, vol. 144, no. 1, pp. 114–126, 2015.

[22] R. E. Petty and J. T. Cacioppo, "The Elaboration Likelihood Model of Persuasion," Advances in Experimental Social Psychology, vol. 19, pp. 123–205, 1986.

[23] M. Eder and H. Sjøvaag, "Artificial intelligence and the dawn of an algorithmic divide," *Frontiers in Communication*, vol. 9, p. 1453251, 2024.

[24] B. Toff and F. M. Simon, “Or they could just not use it? The dilemma of AI disclosure for audience trust in news,” The International Journal of Press/Politics, 2024. https://doi.org/10.1177/19401612241308697

[25] J. Kruger, D. Wirtz, L. Van Boven, and T. W. Altermatt, “The effort heuristic,” Journal of Experimental Social Psychology, vol. 40, no. 1, pp. 91–98, 2004.

[26] iMotions, "Facial Expression Analysis: The Complete Pocket Guide," iMotions, 2024.

[27] L. Kulke, D. Feyerabend and A. Schacht, “A Comparison of the Affectiva iMotions Facial Expression Analysis Software With EMG for Identifying Facial Expressions of Emotion.” Front. Psychol. 11:329. doi: 10.3389/fpsyg.2020.00329, 2020.

Notebook LM, Claude Opus 4.8 and Perplexity were used for literature review, translation and linguistic revision .