Audience reception of news articles made with various levels of automation—and none: Comparing cognitive & emotional impacts

Florian Stalph
Department of Media and Communication
LMU Munich
Germany
florian.stalph@ifkw.lmu.de

Sina Thäsler-Kordonouri
Department of Media and Communication
LMU Munich
Germany
sina.thaesler-kordonouri@ifkw.lmu.de

Neil Thurman
Department of Media and Communication
LMU Munich
Germany
neil.thurman@ifkw.lmu.de

ABSTRACT

Our knowledge about audience perceptions of manually-authored and automated news articles is limited. Although over a dozen studies have been carried out, findings are inconsistent and limited by methodological shortcomings. For example, the experimental stimuli used in some has made isolation of the effects of the actual authorship (automated or manual) difficult. Our study attempts to overcome previous studies' shortcomings to better evaluate audiences' relative evaluations of news articles produced with varying degrees of automation—and none. We conducted a 3 (article source: manually-written, automated, post-edited) × 12 (story topics) between-subjects online survey experiment using a sample (N=4,734) representative of UK online news consumers by age and gender. Each of the 36 treatment groups read a data-driven news article that was either: (1) manually-written by a journalist, (2) automated using a data-driven template, or (3) automated then subsequently post-edited by a journalist. The articles' authorship was not declared. To minimise confounding variables, the articles in each of the 12 story sets shared the same data source, story angle, and geographical focus. Respondents' perceptions were measured using criteria developed in a qualitative group interview study with news consumers. The results show that manually-written articles scored significantly higher on overall liking than automated—but not post-edited—articles. Respondents found manually-written articles to be significantly more comprehensible—both overall and in relation to the numbers they contained—than automated and post-edited articles. Authorship did not have any statistically significant effect on the positive or negative feelings (valence) articles provoked in respondents, or the strength of those feelings (arousal).

The Joint Computation + Journalism European Data & Computational Journalism Conference 2023

CCS CONCEPTS

• Human-centered computing • Human computer interaction (HCI) • Empirical studies in HCI

KEYWORDS

automated journalism, data journalism, survey, audience perception study

ACM Reference format:

Florian Stalph, Sina Thäsler-Kordonouri and Neil Thurman. 2023. Audience reception of news articles made with various levels of automation—and none: Comparing cognitive and emotional impacts. Paper presented at The Joint Computation + Journalism European Data & Computational Journalism Conference 2023. ETH Zurich, Switzerland, 22-24 June 2023.

1 Introduction

News organisations are increasingly deploying automation technologies in news production (e.g., Kotenidis & Veglis, 2021; Dörr, 2016). For example, data mining can augment news discovery, automatically calibrated content can boost personalised experiences, and the automated production of news articles via text generation software can make journalistic content production scalable. This later so-called 'automated journalism' relies primarily on rule-based natural language generation (NLG) systems that use manually created story templates (Graefe & Bohlken, 2020) to transform data into the semantic structure of a readable text. Affected by computational thinking and the technical constraints of these story templates, journalists who automate journalism this way are not writing a story; they are "writing the potential for every eventuality of the story" (Rogers in Diakopoulos, 2019, p. 131) to allow templates to react to variations in datasets. As pointed out by previous studies, the use of such NLG systems could affect the composition of data-driven news articles, including the presence of what Caswell (2019) considers the "essential components of the human craft of journalism" (p. 1149), such as description, background, and anecdotes. Consequently, the increasing use of automated journalism may affect how news consumers perceive news content and necessitates research into the

journalistic performance of stories produced this way. Therefore, this study seeks to gather and compare audience perceptions of articles produced using varying degrees of automation and none and to do so in a way that produces results of unprecedented internal and external validity.

2 Literature Review

Our knowledge about the perceptions of data-driven, automated news articles is limited as few studies have compared the perceptions of news articles written with the full range of automation variants, including automated, post-edited and manually-written articles (to our knowledge, only Wölker & Powell, 2018). Most commonly, studies comparing audience evaluations focus on two variants: automated and manually-written. However, the third variant, where automated articles are post-edited by journalists prior to publication, is increasingly commonly used (see, e.g., Thäsler-Kordonouri & Barling, 2023).

Although over a dozen studies have been conducted on the perception of automated journalism, their findings are inconsistent and often limited by methodological shortcomings.

Automation technologies explored in perception studies range from more human-dependent, template-based applications (e.g. Kolo et al., 2022; Haim & Graefe, 2017) to less human-dependent "modular" NLG systems (Melin et al., 2018, p. 43358) and machine-learning-based applications (Tewari et al., 2021). In several studies, scholars omit to describe how the analysed automation systems operate and refer to their output simply as "software-generated" (Clerwall, 2014, p. 524) or as "generated by algorithms" (Wu, 2020, p. 1015).

Scholars have followed different strategies to find and pair automated and manually-written news articles to be used as experimental stimulus material. In several studies, automatically generated articles were paired with published, manually-written articles based on the same data, or about the same topic or event, as the automated articles (e.g. Kolo, 2022; Wu, 2020; Wölker & Powell, 2018; Haim & Graefe, 2017). Some studies commissioned professional journalists (Jung et al., 2017; Melin et al., 2018) or journalism students (Van der Lee et al., 2018) to write articles to pair with automated stories. In a few cases, articles were paired even though they were written in different journalistic styles, including pairing an automatically generated factual sports report with a manually-written opinion piece (Clerwall, 2014).

In several cases, scholars edited the articles before showing them to readers, for instance, by shortening the manually-written articles to "match the length of the one written by the algorithm" (Jung et al., 2017, p. 295) or by only using the first "400 words" of articles generated by a machine-learning based application (Tewari et al., 2021, p. 12:4). Such artificial interventions might impact the validity of the studies' findings.

Only one study has compared the perceptions of fully automated, manually-written and post-edited articles; doing so with regard to readers' evaluations of source and message credibility (Wölker & Powell, 2018). However, this study also has methodological limitations as one of the post-edited stories was created by the study's authors, which might have reduced the external validity of the stimulus. Furthermore, similar to other studies, the authors only investigated two sets of stimuli. As Jackson and Jacobs (1983) point out, "generalization about a whole category of messages [such as automated or post-edited journalism] requires careful analysis of multiple members of the category" because "any particular message chosen to represent any message category must be assumed to differ from other members of the category in unknown and indefinitely numerous ways" (p. 171). Therefore, we do not believe that the results of a study, such as Wölker and Powell's, that uses just two messages for each message category studied can be generalised to that whole message category.

Our study attempts to overcome previous shortcomings to better evaluate audiences' relative evaluations of news articles produced with varying degrees of automation (and none). Given the apparent absence of literature on how the perception of automated news stories compares with their post-edited offspring, we do not test any hypotheses but rather ask the following research question:

RQ1: How does the perception of automated articles that have been post-edited by human journalists compare with the perception of their automated progenitors and with equivalent, manually-written articles on the same stories and based on the same quantitative data?

3 Methodology

A large-scale 3 (article source: manually-written, automated, post-edited) × 12 (story topics) between-subjects online survey experiment was conducted using a sample (N = 4,792) representative of UK online news consumers by age and gender. The sample of respondents was drawn from various local regions and divided into 36 treatment groups. Each treatment group was exposed to a data-driven news article that had been produced either: (1) manually by a human journalist, (2) using template-based automation, or (3) in a post-edited manner, where a human journalist has further developed the automated article.

Respondents' perceptions were measured using news perception criteria developed in a qualitative pre-study based on group interviews with UK news consumers (N = 31). The 13 criteria cover four domains: *antecedents of perception*, *emotional and cognitive impacts*, *article composition*, and *news and editorial values*. Several of the criteria have not been used in prior research on the perception of data-driven journalism, including that produced with the help of automation. In this paper we focus on one of these domains: the emotional and cognitive impacts the articles have on readers.

3.1 Survey Instrument

3.1.1 Stimulus News Stories. The stimulus material comprises stories sourced from PA Media's Reporters And Data And Robots (RADAR) automated news service and local, regional, and national British online news websites. The sources were chosen purposefully to include a wide range of news outlets and publishers with different geographical foci, backgrounds, funding models, target audiences, and ownership structures. The stimuli cover a range of topics including public health, crime, sport, transport, and social affairs. To eliminate potentially confounding variables, we stripped the articles of bylines and the publishers' logos and branding, showing respondents only the text in basic HTML formatting to ensure readability.

The automated articles (N = 12) were produced by data-driven templates created by data journalists at PA Media's RADAR. The post-edited articles (N = 12) were developed directly from the particular aforementioned automated stories by journalists, who, for example, added quotes from local spokespeople or deleted content that was not deemed relevant to the target audience. We classified articles as post-edited by identifying editorial changes that had been made to the body text of the automated stories. Articles in which only the headline had been changed were not included in the sample of post-edited stories. The final set of articles (N = 12) were purely manually-written (which was confirmed in personal correspondence with the articles' authors) and drew on the same data used in the automated and post-edited versions. The article sets were found via extensive online research.

To minimise confounding variables, the three articles in each of the 12 story sets are based on the same data source(s), feature the same story angle, and cover the same locality.

3.1.2 Perception Criteria and Measures. As we knew from the qualitative interview-based pre-study that the articles could have an *emotional impact* on respondents, we measured *valence*, and levels of *arousal*, and *liking*. Respondents indicated the intensity and direction of emotional arousal they experienced when reading an article (see Kuppens et al., 2013). Additionally, respondents reported their overall 'liking' of the news article (see Sundar, 1999). These variables were measured on a continuous scale modelled after the affective slider, a self-reporting tool for the quick assessment of pleasure and arousal (Betella & Verschure, 2016).

Aside from emotional affect, our qualitative interview-based pre-study showed that articles can also have a *cognitive impact* on readers. Therefore, we asked respondents about the *overall readability and comprehensibility* of the article they read. In addition, we asked about respondents' *comprehension of numbers* in the article.

3.1.3 Instrument Pre-Test. The questionnaire was pre-tested in October 2022 after its initial development but before its full-scale field administration. We used developmental expert reviewing and cognitive interviews (N = 10) with respondents; involving think-aloud complemented with verbal probing procedures (Willis, 2016). Our goal was to identify and repair problematic measures, questions and concepts, and issues with the usability of the survey. To generate feedback, we administered a prototype of the survey that contained nine different articles. Three of these articles were manually-written, three automated, and three post-edited to ensure that the pre-test covered articles produced with all potential degrees of automation, including none. To test variations in the responses related to individual reactions to story topics, the articles covered nine different topics, including energy costs, e-scooter casualties, and drug deaths. Each article was sourced from Birmingham news outlets and we only recruited pre-test candidates who lived in the Birmingham area. By doing so, we took special care to ensure that our respondents were exposed only to sets of stories that were relevant to their geographic interest. Based on the feedback collected from the pre-test interviews, we modified questions and repaired survey defects (Willis, 2016).

Before the survey's broader distribution we deployed a soft launch. Approximately 100 respondents completed the survey allowing us to test its technical functionality and measures, check output data structure and validity, and to gather additional feedback via a free-text question.

3.1.4 Survey Administration and Data collection. The survey was fielded by YouGov to their own proprietary online panel between 26 January and 1 March 2023. To be eligible for participation in the survey, respondents were pre-screened to ensure that they were aged 18 or older, used online news at least once a month, and were resident in one of the selected news organisations' catchment areas to ensure that the article was relevant to where they lived. Each experimental treatment group (automated, manual, and post-edited) comprised at least 100 participants, with quotas set on age and gender, reflective of local populations.

To make sure that the recruited participants read the stimulus article, only respondents who passed an attention check were included in the final sample (Ruble, 2017).

3.1.4 Sample Description. The overall target population for this survey was monthly UK news consumers ages 18 and older, who were resident in regions and cities covered by catchment areas of those local news organisations we drew our stimulus news articles from.

The sample comprises N = 4,734 respondents with a mean age of M = 50.66 years (SD = 15.77) and a gender split of 55 percent women (N = 2602) and 45 percent men (N = 2132). Each experiment within each area required at least 100 respondents per

treatment (automated, manual, and post-edited), amounting to at least 300 respondents per experiment due to oversampling.

4 Results

4.1 Differences in Levels of Liking of Automated, Manually-written, and Post-edited Articles

We conducted a one-way ANOVA to assess the effects of level of automation on level of liking. We removed outliers, identified according to inspection with a box-plot, decreasing the sample size for this test to N = 4,595. There was homogeneity of variance (Levene's test, $p > .05$) for level of liking for each group.

The level of liking differed significantly for the different levels of automation, F(2, 4592) = 7.43, $p < .001$, $\eta^2 = .003$. A Tukey post-hoc analysis (-1.51, 95%-CI[-2.44, -.59]) revealed a significant difference ($p < .001$) between levels of liking for automated (M = 3.49, SD = 11.21) and manually-written articles (M = 5.01, SD = 11.55). However, there were no significant differences between levels of liking for automated and post-edited articles (M = 4.11, SD = 11.01) or manually-written and post-edited articles.

4.2 Differences in Emotional Impact of Automated, Human, and Post-edited Articles

We conducted a one-way ANOVA to assess the effects of level of automation on arousal.

We removed outliers, identified according to inspection with a box-plot. Homogeneity of variances was asserted using Levene's Test which showed that equal variances could not be assumed (p = .003).

The highest levels of arousal across the respondents (N = 4,710) could be found after they had read one of the automated articles (M = 5.28, SD = 13.36). Arousal was weaker after respondents had read a post-edited (M = 5.17, SD = 14.03) or manually-written article (M = 5.07, SD = 14.42). However, arousal did not differ significantly for the different levels of automation, Welch's $F(2, 3074.60)$ = .100, $p > .05$.

We conducted another one-way ANOVA to assess the effects of level of automation on valence. We removed outliers, identified according to inspection with a box-plot, decreasing the sample size for this test to N = 4,680. Homogeneity of variances was asserted using Levene's Test which showed that equal variances could not be assumed ($p < .001$).

The automated articles triggered the highest negative valence (M = -6.83, SD = 11.93), followed by post-edited articles (M = -6.26, SD = 12.22), and manually-written articles (M = -5.95, SD = 13.67). However, valence did not differ significantly for the different levels of automation, Welch's $F(2, 3070.81) = 2.04$, p = .131.

4.3 Differences in Cognitive Impact of Automated, Manually-written, and Post-edited Articles

To explore differences in respondents' evaluations of the overall comprehensibility of the articles, we conducted a one-way ANOVA comparing articles with different authorships. We excluded extreme values for the variable per group, identified according to inspection with a box-plot, decreasing the sample size for this test to N = 4,640. Homogeneity of variances was asserted using Levene's Test, which showed that equal variances could not be assumed ($p < .001$). The overall comprehensibility of the articles differed significantly for the different levels of automation, Welch's $F(2, 2975.68) = 48.03$, $p < .001$, $\eta^2 = .028$.

Games-Howell post-hoc analysis revealed a significant difference ($p < .001$) in overall comprehensibility between manually-written and automated articles as well as between manually-written and post-edited articles. The mean value for how comprehensible respondents thought articles were overall was higher for manually-written articles than for automated articles (3.79, 95%-CI[2.80, 4.79]), and also higher for manually-written than post-edited articles (3.23, 95%-CI[2.19, 4.27]).

To explore differences in respondents' comprehension of numbers across differently produced articles, we conducted a one-way ANOVA. We excluded extreme values for the variable per group, identified according to inspection with a box-plot, decreasing the sample size for this test to N = 4,710. Homogeneity of variances was asserted using Levene's Test which showed that equal variances could not be assumed ($p < .001$). Respondents' comprehension of numbers in the articles differed significantly for the different levels of automation, Welch's $F(2, 3014.14) = 72.97$, $p < .001$, $\eta^2 = .028$.

Games-Howell post-hoc analysis revealed a significant difference ($p < .001$) in the comprehension of numbers between manually-written and automated articles and between manually-written and post-edited articles. The mean value for how comprehensible respondents thought numbers in the articles were was higher for manually-written articles than for automated articles (5.05, 95%-CI[3.97, 6.13]), and also higher for manually-written than post-edited articles (4.32, 95%-CI[3.19, 5.46]).

4 Discussion

This study investigated the differences in news consumers' liking of automated, manually-written, and post-edited news articles, and the emotional and cognitive impacts those different articles types had on them. The results show some significant differences in the level of liking for the three types of articles, with manually-written articles being significantly more liked than automated articles. However, there were no significant differences between the liking of post-edited and manually-written articles or between post-edited

and automated articles. Regarding emotional impact, although automated articles triggered the highest level of arousal, followed by post-edited and manually-written articles, the differences were not statistically significant. In terms of valence, although automated articles triggered the highest level of negative valence, followed by post-edited and manually-written articles, again these differences were not statistically significant. In terms of cognitive impact, there were significant differences between the article types. Manually-written articles were significantly more comprehensible than automated and post-edited articles. The comprehension of numbers in articles was also significantly higher for manually-written articles than for automated and post-edited articles.

The findings of this study have implications for the design and production of news articles. The results suggest that manually-written articles are generally more appealing and comprehensible than fully-automated articles, indicating the importance of maintaining human involvement in the production of news content. However, the study also showed that post-edited articles were not significantly less liked than manually-written articles, suggesting that automated content may be more appealing if it is post-edited by humans.

The study's findings also suggest that news consumers found automated and post-edited articles less comprehensible than their manually-written counterparts, particularly when it comes to understanding numerical information. News outlets should consider providing additional context and explanations to help readers understand complex information and data in news articles.

Overall, the study highlights the importance of finding the right balance between human involvement and automation to ensure that news content is both appealing and comprehensible to the audience.

ACKNOWLEDGMENTS

This work was supported by Volkswagen Foundation: [Grant Number A110823/88171].

REFERENCES

[1] Alberto Betella & Paul F. M. J. Verschure, 2016. The Affective Slider: A Digital Self-Assessment Scale for the Measurement of Human Emotions. *PLoS ONE 11*(2), e0148037. DOI:https://doi.org/10.1371/journal.pone.0148037

[2] Christer Clerwall, 2014. Enter the Robot Journalist. *Journalism Practice 8*(5), 519-531. DOI:http://dx.doi.org/10.1080/17512786.2014.883116

[3] Andreas Graefe & Nina Bohlken, 2020. Automated Journalism: A Meta-Analysis of Readers' Perceptions of Human-Written in Comparison to Automated News. *Media and Communication 8*(3), 50-59. DOI:https://doi.org/10.17645/mac.v8i3.3019

[4] Mario Haim & Andreas Graefe, 2017. Automated News. *Digital Journalism*, 1-16. DOI:http://dx.doi.org/10.1080/21670811.2017.1345643

[5] Sally Jackson & Scott Jacobs, 1983. *Human Communication Research 9*(2), 169-191.

[6] Jaemin Jung, Haeyeop Song, Youngju Kim, Hyunsuk Im & Sewook Oh, 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. DOI:https://doi.org/10.1016/j.chb.2017.02.022

[7] Castulus Kolo, Joschka Mütterlein & Sarah Anna Schmid, 2022. Believing Journalists, AI, or Fake News: The Role of Trust in Media. *55th Hawaii International Conference on System Sciences*. DOI:https://doi.org/10.24251/HICSS.2022.393

[8] Peter Kuppens, Francis Tuerlinckx, James A. Russell & Lisa Feldman Barrett, 2013. The relation between valence and arousal in subjective experience. *Psychological Bulletin 139*(4), 917–940. DOI:https://doi.org/10.1037/a0030811

[9] Magnus Melin, Asta Bäck, Caj Södergård, Myriam D. Munezero, Leo J. Leppänen & Hannu Toivonen, 2018. No Landslide for the Human Journalist - An Empirical Study of Computer-Generated Election News in Finland. *IEEE Access 8*, 43356-43367. DOI:https://doi.org/10.1109/ACCESS.2018.2861987

[10] Shubhra Tewari, Renos Zabounidis, Ammina Kothari, Rexnold Bailey & Cecilia Ovesdotter Alm, 2021. Perceptions of Human and Machine-Generated Articles. *Digital Threats: Research and Practice 2*(2), 1-16. DOI:https://doi.org/10.1145/3428158

[11] Chris van der Lee, Bart Verduijn, Emiel Krahmer & Sander Wubben, 2018. Evaluating the text quality, human likeness and tailoring component of PASS: A Dutch data-to-text system for soccer. In *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA, August 20-26, 2018, 962–972.

[12] Sina Thäsler-Kordonouri & Kurt Barling, 2023. Automated Journalism in UK Local Newsrooms: Attitudes, Integration, Impact. *Journalism Practice*, 1-18. DOI:https://doi.org/10.1080/17512786.2023.2184413

[13] Anja Wölker & Thomas E. Powell, 2018. Algorithms in the newsroom? News readers' perceived credibility and selection of automated journalism. *Journalism*, 1-18. DOI:https://doi.org/10.1177/1464884918757072

[14] Yanfang Wu, 2019. Is Automated Journalistic Writing Less Biased? An Experimental Test of Auto-Written and Human-Written News Stories. *Journalism Practice,* 1-21. DOI:https://doi.org/10.1080/17512786.2019.1682940