Article

Readers' perception of computer-generated news: Credibility, expertise, and readability

Andreas Graefe
Columbia University, USA; LMU Munich, Germany; Macromedia University, Germany

Mario Haim
LMU Munich, Germany

Bastian Haarmann
Fraunhofer Institute for Communication, Information Processing and Ergonomics, Germany

Hans-Bernd Brosius
LMU Munich, Germany

Abstract
We conducted an online experiment to study people's perception of automated computer-written news. Using a 2 × 2 × 2 design, we varied the article topic (sports, finance; within-subjects) and both the articles' actual and declared source (human-written, computer-written; between-subjects). Nine hundred eighty-six subjects rated two articles on credibility, readability, and journalistic expertise. Varying the declared source had small but consistent effects: subjects rated articles declared as human written always more favorably, regardless of the actual source. Varying the actual source had larger effects: subjects rated computer-written articles as more credible and higher in journalistic expertise but less readable. Across topics, subjects' perceptions did not differ. The results provide conservative estimates for the favorability of computer-written news, which will further increase over time and endorse prior calls for establishing ethics of computer-written news.

Corresponding author:
Andreas Graefe, Department of Communication Studies and Media Research, LMU Munich, Oettingenstraße 67, 80538 München, Germany.
Email: a.graefe@lmu.de

Introduction

Although computers have long assisted journalists with their daily work (e.g. in researching facts and analyzing data), journalists have remained the sole creators of news. This division of labor is, however, currently changing. Advances in the fields of information technology, linguistics, and natural language generation have made it possible for algorithms to autonomously write news stories from data that are stored in a structured and machine-readable form. This development is sometimes referred to as automated or robot journalism (Carlson, 2015; Clerwall, 2014; Lemelshtrich Latar, 2015; Napoli, 2014) and is part of a larger trend known as computational journalism, which describes an increasing influence of computation and data on journalism (Anderson, 2013; Cohen et al., 2011; Lewis and Westlund, 2015).

For publishers, the most obvious benefit of computer-written news is an economic one: computers are able to generate news at a much larger scale, and thus at a lower cost, than human journalists (van Dalen, 2012). Today, companies such as Narrative Science and Automated Insights provide algorithms that generate millions of articles on topics such as sports, finance, and marketing (Ulanoff, 2014), and publishers have already begun to use computer-written stories in their news coverage. Forbes, for example, has been using this technology since 2012 to report on company earnings (Levy, 2012). In addition to economic benefits, companies are seeking to take advantage of the speed with which computers can generate news. A recent example is the Associated Press, which published a computer-written report on Apple's quarterly earnings only minutes after the company released its figures in January 2015 (AP, 2015).

The computer-written news industry is expected to grow quickly. Kristian Hammond, co-founder of Narrative Science, predicted that computers will write more than 90 percent of news by 2025 (Levy, 2012). While this number is certainly debatable, automated journalism is likely to disrupt news writing in the years to come. The expected growth in computer-written content will result in additional news that is not available today, since computers will report on small-scale events that journalists are unwilling to cover or for which publishers are unwilling to hire journalists. Besides simply increasing the quantity of news, the growth of computer-written news is also expected to affect how journalists create and how consumers perceive news (Graefe, 2016).

Scholars have only just begun to study the implications of this development. In particular, researchers have looked at how the widespread adoption of computer-written news may potentially impact the traditional role of journalists and the quality of news coverage in general. While some argue from a theoretical point of view (Lemelshtrich Latar, 2015; Napoli, 2014), two studies analyzed journalists' news coverage of Automated Insights (van Dalen, 2012) and Narrative Science (Carlson, 2015), two leading companies in automated news generation.

Regarding the potential impact on journalists' traditional roles, some take a positive view in that computer-written news will aid journalists in their daily work. For example, journalists could delegate routine tasks (e.g. reporting company earnings or recapping sport events) to algorithms. Computer-written stories could also provide a first draft that covers basic facts, which journalists would then verify or enrich with more in-depth analyses and interpretation. As a result, so the theory goes, journalists would have more time available for higher value and labor-intensive tasks such as investigative reporting (van Dalen, 2012). An example is crime reporting by the L.A. Times Homicide Report, in which an algorithm provides basic facts, such as the date, location, time, age, gender, race, and jurisdiction of a homicide. In a second step, journalists can add to a story by providing details about the victim's life and family (Young and Hermida, 2015).

Those with a more skeptical view suggest that the increasing adoption of computer-written news will put pressure on journalists. In particular, those who currently perform routine tasks in areas with highly structured and rich databases (e.g. sports, finance, weather) will likely be unable to compete with automatic data collection and writing (Carlson, 2015).

Apart from publishers' economic considerations, the question of whether algorithms will augment or supplant journalists will likely depend on an article's topic, its purpose, and the consumers' expectations regarding the quality of the content. That is, while algorithms can provide answers to clearly formulated questions by analyzing given data, they cannot raise questions on their own or provide opinions on important social issues or proposed policy changes, at least not yet (Lemelshtrich Latar, 2015). Skeptics also predict that news consumers would dislike reading computer-written stories. The reason is that algorithms are limited in understanding and producing nuances of human natural language such as humor, sarcasm, and metaphors. As a result, computer-written stories tend to sound technical and boring (Lemelshtrich Latar, 2015).

Proponents, however, argue that the ability of algorithms to generate natural human language will improve, which will make the content more appealing to news consumers. More importantly, computer-written news could potentially increase the quality and objectivity of news coverage. One argument is that computers never get tired. Thus, algorithms are less error-prone, as they do not make mistakes such as overlooking facts. Another argument is that algorithms strictly follow predefined rules for converting data to text and are thus capable of an unbiased account of facts. The latter argument is based on the assumption that the underlying data are complete and correct and, more importantly, the algorithms are programmed correctly and without bias. This view, however, is rather optimistic: like any other model, algorithms for generating computer-written news rely upon data and assumptions, which both are subject to biases and errors (Lazer et al., 2014). As a result, algorithms could produce outcomes that were unexpected and unintended (Diakopoulos, 2015). First, the underlying data may be wrong, biased, and incomplete. Second, the assumptions built into the algorithms may be wrong or could reflect the conscious or unconscious biases of those who developed or commissioned them (Lemelshtrich Latar, 2015).

Given that developers are unlikely to fully disclose their algorithms, it remains unclear how to evaluate the quality of the algorithms that generate computer-written articles. A promising yet complex approach might be reverse engineering, which aims at decoding

Table 1. Experimental studies of perception of computer-generated news available to date.

		Article source *declared as* written/generated by		
		Journalist	Algorithm	Not declared
Actual article source	Human	(1)	(2)	(5) (Clerwall, 2014)
	Computer	(3) (van der Kaa and Krahmer, 2014)	(4) (van der Kaa and Krahmer, 2014)	(6) (Clerwall, 2014)

Treatments analyzed in this study: (1), (2), (3), and (4).

an algorithm's set of rules by varying certain input parameters and assessing the effects on the outcome (Diakopoulos, 2015). An alternative approach is to analyze how news consumers *perceive* the quality of computer-written news in relation to human-written news. This approach, which can be regarded as a Turing test of journalism, is the route taken in this study. In particular, we build on – and extend – prior work on how recipients perceive computer-written news. In the remainder of this article, we review previous studies and report new evidence from an online experiment conducted in Germany with 986 participants.

Prior evidence

Prior evidence on the perception of computer-written articles is limited. We are aware of only two studies, which differ in their experimental designs (see Table 1).

Clerwall (2014) analyzed differences in perceived quality, measured as credibility and readability, of news articles. He presented 46 Swedish undergraduates in media and communication studies with one of two versions of an article on an American Football game. The article was generated either by a journalist or by a computer but the experiment participants were not given any information about the source (i.e. treatments 5 and 6 in Table 1). This setting thus reflected a situation in which publishers do not byline news stories, which is not uncommon for wire stories (e.g. Associated Press) and computer-written news (Ulanoff, 2014). The articles were written in English, contained no pictures, and were approximately of the same length. After reading and assessing the article's credibility and readability, participants also had to guess whether the article was written by a journalist or generated by a computer.

Overall, differences were small, which is not surprising given the sample size, and participants were unable to correctly identify the article source. However, the direction of effects revealed that the computer-written articles received higher ratings on credibility, whereas the articles written by the journalist scored higher on readability.

The results might surprise. Communication students – who one could expect to have a higher level of media literacy than average news consumers – were unable to distinguish between articles generated by a computer and those written by a journalist, and even favored the computer-written article in terms of credibility.

Often, however, news consumers are informed whether an article was written by a journalist or created by a computer. Forbes, for example, adds the byline 'by Narrative Science' to computer-written articles. Similarly, Associated Press reveals if a story was computer-written (e.g. 'This story was generated by Automated Insights using data from Zacks Investment Research'). It remains unclear, though, whether consumers understand the meaning of such bylines.

Thus, one might ask whether perception changes if consumers know – or think they know – that a computer generated an article. This question was addressed by van der Kaa and Krahmer (2014), who studied the perceived credibility, measured as trustworthiness and journalistic expertise, of computer-written articles. The authors presented 232 native Dutch speakers (168 regular news consumers and 64 journalists) with a computer-written article that either reported the results of a sports event (i.e. a soccer game) or financial news (i.e. stock prices). The articles were written in Dutch and contained no pictures. The authors then manipulated the byline of the article, which was either correctly declared as 'written by a computer' or wrongly declared as 'written by a journalist' (i.e. cases 3 and 4 in Table 1). Among regular news consumers, differences in perceived expertise and trustworthiness were small, although articles declared as written by the computer received slightly higher ratings on both dimensions than those declared as written by the journalist. In comparison, the 64 journalist participants assigned higher ratings of trustworthiness to articles that were declared as written by the journalist; no differences were found for journalists' perceptions of expertise. The study also revealed some differences regarding the story topic: while there were no significant differences on expertise, the soccer articles scored lower on trustworthiness than the finance articles.

In sum, the results conformed to those of Clerwall (2014). There was little difference in news consumers' perceived credibility of articles, regardless of whether the articles were labeled as written by a computer or by a journalist. When discussing potential reasons for the small differences, van der Kaa and Krahmer (2014) suggested that news consumers' initial – perhaps subconscious – expectations might influence the results in favor of computer-written articles. In particular, subjects might have high expectations when reading an article (declared as) written by a journalist and low expectations when reading an article (declared as) generated by a computer. If subjects are positively surprised by the quality of a computer-written article (and several of their experiment participants reported that they were), they might assign higher ratings. In contrast, if subjects' expectations in the quality of a human-written article are not fulfilled, they might assign lower ratings.

This study addresses this question by building on – and extending – the experimental design of the two previous studies. In particular, we mimic the design of van der Kaa and Krahmer (2014) by varying the declared article source. However, we also vary the actual source as either human-written or computer-written. This variation allows us to, first, analyze whether news consumers' perceptions are indeed influenced by their initial expectations regarding the (declared) article source. If so, they should rate human-written articles higher if they are wrongly declared as computer-written. More importantly, second, this experimental design enables us to study differences in perceptions of articles that are written by journalists and correctly declared as such (case 1 in Table 1) and articles that are computer-written and correctly declared as such (case 4 in Table 1). This

question, which has not been analyzed in prior research, is important as it has decision-making implications for publishers, who need to decide whether to hire human journalists or have algorithms produce articles on certain topics. Finally, our setting allows for testing whether the previous findings hold in a different country (i.e. Germany).

Method

This section describes our measures of news perception, the experimental design, the online questionnaire, the participants, and the stimulus material. For additional information, such as the complete questionnaire and the full article texts see the online appendix (Graefe et al., 2016).

Measuring news quality

Quality of news is a fuzzy concept and difficult to measure as it means different things to different people. For example, some people might assess an article's quality based on the excellence of writing (e.g. the use of stylistic devices), whereas others might focus on whether the article is well researched or fits their view of the world. Thus, perceived quality does not necessarily relate to objectively measured quality (Urban and Schweiger, 2014). Nevertheless, analyzing news consumers' perceptions of different aspects of quality, such as credibility, is common in studies of news quality.

Since Hovland et al. (1953), numerous scholars have evaluated the quality of news. Building upon seminal work by Meyer (1988) and West (1994), scholars have distinguished between the credibility of a news article's source (e.g. Flanagin and Metzger, 2008; Metzger et al., 2010), its message (e.g. Westerman et al., 2014), and its distributing medium (e.g. Golan, 2010; Johnson and Kaye, 2004). Thereby, researchers commonly use factor analysis on a variety of items to identify different dimensions of news quality. For example, Sundar (1999) presented people with news articles and asked for their open-ended quality evaluations, which he condensed to 21 different items. He then validated the 21 items by obtaining ratings for different articles from different subjects. This procedure identified four central factors that people consider when evaluating news content: credibility, readability, quality, and representativeness.

More generally, researchers have shown that news quality builds upon multiple dimensions and that its adequate measure requires using items that match the specific type (e.g. written or audio news) and topic (e.g. sports or finance) of the news (Kohring and Matthes, 2007). As a result, the literature lacks specific guidelines for how to measure news quality and, not surprisingly then, different researchers have used different scales. Looking at the two previous studies on perception of computer-written news, we can see such differences at work. Clerwall (2014) measured quality by obtaining readers' perceptions of credibility (i.e. whether they found the articles informative, trustworthy, objective, and descriptive) and readability (i.e. whether they found the articles pleasant to read, clear, well written, coherent, and not boring). Since his experimental design did not reveal the article source, he essentially measured message credibility. In comparison, van der Kaa and Krahmer (2014) revealed an (either true or false) article source and thus measured message and source credibility by obtaining ratings for trustworthiness (i.e.

reliability, honesty, accuracy, and fact-based) and journalistic expertise (i.e. expertise, intelligence, authority).

In order to develop a measure of content perception for this study, we followed an approach similar to Sundar (1999). We conducted a pretest with 40 participants, who were asked to rate a computer-written article (either on soccer or finance) on 21 items, using a 7-point scale from 'I completely agree' to 'I completely disagree'. These included 17 items from Sundar (1999) – except for indifferent items and items that relate to 'representativeness', as this dimension does not cover single news items but rather selections of news (e.g. a front page) – as well as four items (trustworthy, complete, descriptive, fact-based) previously used by van der Kaa and Krahmer (2014) and Clerwall (2014). In addition, respondents could suggest new items. However, because only four respondents made suggestions and there was no coherence in their responses, we dismissed the open-ended answers. Exploratory factor analysis yielded three dimensions, each of which is based on four items that capture perceptions of credibility (accurate, trustworthy, fair, and reliable), readability (entertaining, interesting, vivid, and well written), and journalistic expertise (coherent, concise, comprehensive, and descriptive).

Stimulus material

In a similar way to van der Kaa and Krahmer (2014), we selected articles from the domains of sports and finance, which are representative of the current use of computer-written news. The reason is that, for both topics, reporting is commonly fact-based, and the underlying data (e.g. game statistics, historical and current stock prices, etc.) are widely available in a structured format. The Fraunhofer Institute for Communication, Information Processing and Ergonomics provided us with the computer-written articles, one each for sports and finance. Their software is currently being used by Finanzen100. de, a German website that publishes financial reports and which is part of the Focus Online group. For more information about the technology, see Haarmann and Sikorski (2015). The articles were written in German and referred to events from the previous week. In order to avoid high involvement, the soccer article reported on a game from Germany's second division. The finance article reported on a German car producer's share performance.

The human-written stories covered the same events and were obtained from popular German websites for sports (i.e. sport1.de) and finance (i.e. deraktionaer.de, the online edition of a weekly financial magazine). To assure external validity, none of the articles were edited or shortened, except for removing pictures. We also presented the different versions to nine master students in communication science and asked them to identify the computer-written articles; the students were correct about 50 percent of the time, which reflects random guessing.

Experimental design

The experimental $2 \times 2 \times 2$ design draws on ideas from both previous studies. First, in a similar way to Clerwall (2014), we varied the *actual* article source (i.e. whether the article was actually written by journalist or was generated by a computer). Second, like van

der Kaa and Krahmer (2014), we varied the *declared* article source by adding a byline that labeled the article as computer-written or human-written. As shown in Table 1, this between-subjects design results in four treatments:

1. Human-written articles correctly declared as such;
2. Human-written articles wrongly declared as computer written;
3. Computer-written articles wrongly declared as human-written;
4. Computer-written articles correctly declared as such.

Each participant was presented one article of each topic, in randomized order (within-subjects design). That is, participants either saw a soccer article first, followed by a finance article, or vice versa. Moreover, each participant saw one article declared as written by an algorithm and one article declared as written by a journalist; the actual article source was randomized. In other words, if a participant was assigned to either scenario (1) or (3) for the first article, then she was assigned to either scenario (2) or (4) for the second article, and vice versa.

Questionnaire

In an online questionnaire, participants were first asked about media usage patterns, journalistic experience, and interest in different topics. Then, participants entered the experimental setting, where they had to rate both articles on the 12 measures described above, using a 5-point scale ranging from 'I completely agree' to 'I completely disagree'. Finally, participants were asked for socio-demographic information. Participants spent, on average, 8.5 minutes (SD = 3.2) completing the questionnaire.

Subjects

Participants were recruited through the SoSci Panel, a noncommercial online access convenience panel, whose approximately 90,000 active members voluntarily participate in scientific surveys. The panel has two major advantages compared to traditional student samples. First, it allows for the recruitment of a large number of participants and thus addresses the small sample problem of previous studies. Second, the resulting samples are more heterogeneous than student samples regarding age, geography, professional background, and personal interests. However, the panel members are not representative of the German-speaking population. In particular, they are better educated than the general public. In addition, given the exclusive use of online surveys, the panel members have a generally high affinity for the Internet, which makes them a suitable target group for this study. For more information on the SoSci Panel, see Leiner (2014).

A total of 1107 subjects participated in the study in December 2014. After removing incomplete questionnaires, 986 subjects (53% female) remained. The average age was 38 years, 55 percent had at least a university degree. There were no statistically significant differences across the experimental groups in terms of age, gender, media usage patterns, prior journalistic experience, and interest in sports and finance.
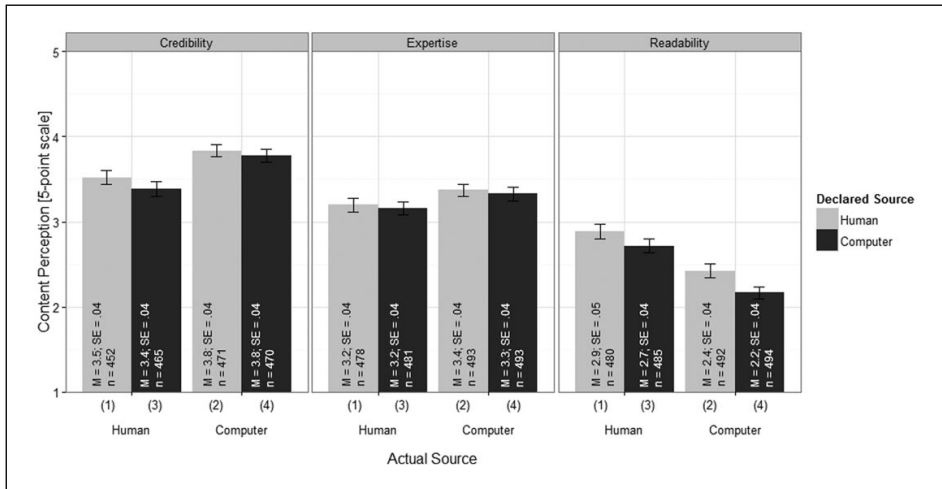
Figure 1. Mean ratings per group.
Values represent means.
Error bars indicate 95 percent confidence intervals.
Treatments (1) to (4) as shown in Table 1.

Results

Cronbach's α suggests that our measures of the three dependent constructs were reliable (credibility: $\alpha = .83$; readability: $\alpha = .85$; expertise: $\alpha = .76$). Figure 1 shows the results over both topics. The columns show the mean ratings per construct and group. The error bars depict 95-percent confidence intervals and thus indicate statistical significance; Appendix 1 shows results from a multivariate analysis of variance. All data and calculations are publicly available in the online appendix (Graefe et al., 2016).

Effect of the topic

Finance articles scored between 0.1 and 0.5 points (on the 5-point scale) lower than sports articles on each of the three dimensions. The direction of the effects, however, was identical across both topics. We therefore merged the data in order to simplify the presentation of our findings. Results per topic are available in Appendix 2.

Effect of the declared source

The effect of the declared source was consistent across the three quality measures (i.e. credibility, expertise, and readability). That is, regardless of the actual source, articles were always rated higher if declared as written by a journalist. In all but one case, however, differences were rather small and, thus, not statistically significant (i.e. the confidence intervals overlapped). The one exception was the computer-written articles, which

were rated substantially higher in terms of readability if they were (wrongly) declared as written by a journalist.

Effect of the actual source

The actual source's effect on people's perception differed across the three constructs. Regardless of the declared source, the computer-written articles were rated as more credible and higher in terms of expertise than the human-written articles. For example, the correctly declared computer-written articles received a mean rating of 3.8 on the 5-point scale, which is a quarter of a point (or 7%) higher than the corresponding ratings for correctly declared human-written articles.

In contrast, the results for readability showed the opposite effect: human-written articles scored significantly higher than those written by the computer. Differences were particularly large if the article source was declared correctly: the mean rating for the human-written article (2.9) was 0.7 points (or 34%) higher than the rating for the computer-written article (2.2).

Discussion

Our findings corroborate those of two previous studies (Clerwall, 2014; van der Kaa and Krahmer, 2014), which used different experimental designs with different measures of news quality, were conducted in different countries, and were based on substantially smaller samples of participants. First, computer-written news tends to be rated higher than human-written news in terms of credibility. Second, news consumers get more pleasure out of reading human-written as opposed to computer-written content. Third, differences in terms of perceived credibility and expertise tend to be small. A possible explanation for the small differences is that algorithms strictly follow standard conventions of news writing and, as a result, computer-written stories reflect these conventions. Given that a major portion of news writing is a simple recitation of facts and often lacks sophisticated storytelling and narration, it is not surprising that recipients rated both article sources as rather credible and expert. Interestingly, however, the recipients did not like reading either type of article very much. Although the human-written ones were rated as clearly more readable than computer-written stories, their average rating was still below the mid-point of the 5-point scale. One explanation for low readability ratings might be that sports and finance are boring subjects for many people. Another explanation might be that the results indicate a general dissatisfaction with news writing for such topics.

For many topics, and in particular routine tasks, publishers will increasingly have the opportunity to rely on services that create computer-written news, rather than hiring a journalist to write a story. Thus, an important comparison, which has not been analyzed in prior research, is to compare news consumers' perception of articles that are written by journalists and correctly declared as such, with articles that are computer-written and correctly declared as such. Our results show that subjects rate computer-written articles slightly higher in terms of credibility and journalistic expertise, whereas human-written articles score significantly higher in terms of readability. Given the current state of the

technology and apart from economic considerations, publishers thus face a trade-off between credibility and readability when deciding between computer- and human-written stories. That said, the readability of computer-written news is likely to further improve over time, as computer linguists are constantly enhancing their algorithms' ability to analyze large data sets and to generate natural human language such as humor or poetry (Gonçalo Oliveira and Cardoso, 2015; Petrovic and Matthews, 2013). In comparison, it is rather unlikely that the quality of human journalists will equally improve – at least not at the same pace. In the short term, we would thus expect follow-up studies to find even stronger effects in favor of computer-written content. However, such effects may not necessarily persist in the long term. It may well be that after readers' initial excitement with the new technology, algorithmic news that builds on a static set of rules might 'get old', in particular, if used at a large scale. If so, readers may be drawn toward fresh and creative human writing styles again, which may create new opportunities for journalists. Future research should track how the quality of both human- and computer-written news will evolve and how people's expectations toward and perceptions of such content may change over time.

The results further show that articles are consistently perceived more favorably if they are declared as written by a human journalist, regardless of the actual source. This finding has two important implications. First, the results address the question raised by van der Kaa and Krahmer (2014), who suggested that consumers' initial expectations regarding the quality of the declared article source might influence their perceptions of quality. In particular, they argued that consumers might have low expectations for computer-written articles and might thus be positively surprised by their quality, which, in turn, would lead to higher ratings. If this is true, then human-written articles should score higher when they are declared as computer-written. Our results suggest that this is not the case. In fact, the effects pointed in the opposite direction: human-written articles that were wrongly declared as computer-written were perceived as less favorable than the same articles correctly declared as written by a journalist. Second, although differences in effect sizes were small, the results might tempt publishers to assign human names to computer-written articles. The results therefore endorse prior calls for establishing ethics of robot journalism (e.g. Diakopoulos, 2015; Lemelshtrich Latar, 2015). Publishers, for example, should commit to faithfully revealing who created an article.

In sum, the available evidence suggests that the quality of computer-written news is competitive with that of human journalists for routine tasks for which there are well-structured, machine-readable, and reliable data. In such situations, news-generating algorithms excel by quickly extracting information from data, weighting information by importance, generating news narratives, and varying writing styles. Popular applications of computer-written news currently include data-heavy domains such as weather forecasting, sports news, traffic reporting, financial analysis, earthquake warnings, and crime reporting (Young and Hermida, 2015). For such routine tasks, journalists may face the danger of eventually being replaced by automated journalism (Graefe, 2016).

It is important to note that our results cannot be generalized to topics that are not solely fact-based and for which journalists contribute value by providing interpretation, reasoning, and opinion, for example, when it comes to discussing social and political issues. Currently, computer-written stories for such complex problems are not yet

available. However, we expect that the quality of computer-written news will continue to improve, which might enable algorithms to generate journalistic output other than simply reciting facts. Already today, algorithms rely on predefined rules to obtain additional insights from the data. For example, for previews of soccer games, the algorithms take into account the teams' historical record (against each other) or which players will need to avoid a booking in order to not be suspended the following game.

Automated journalism may comprise a major share of news writing in the future. Due to the ever-increasing availability of data, algorithms will be able to cover events where currently limited news is available (Carlson, 2015). In addition, algorithms will likely be able to write stories that suit individual readers' interests, political views, and education levels. This development raises questions about possible implications about the future of journalism and its relationship to the democratic process. Given the idea that journalism (among others) should enable citizens to act on politically well-informed grounds, automated journalism could both hinder and foster these ideals. On the positive side, the ability to personalize news may make it possible to attract a broader audience and thus increase the number of politically informed people. Furthermore, if automation of routine tasks will indeed free up resources, journalists might have more time for in-depth analysis, which could improve news quality. On the negative side, an increasing quantity of available news will further increase people's burden to find news that is most relevant to them. Given news consumers' limited time, available offers need to be customized to individual needs, for instance by increasingly relying on personalized news aggregators such as Google News or relevance-sorting algorithms such as Facebook's news stream. Such increasing personalization could lead to fragmentation or 'filter bubble' (Pariser, 2011) effects within society. The concern is that personalization leads individuals to consume more and more of the same information. As a result, people are unlikely to consume information that challenges their views or contradicts their interests, which may carry risks for the formation of public opinion in a democratic society.

An increasing use of computer-written news also raises more general questions as to whether we can and should trust algorithms as a mechanism to provide checks and balances, to identify important issues, and to establishing a common agenda for the democratic process of public opinion formation. For instance, algorithms could analyze publicly available information such as annual reports or speech protocols in order to provide insights on political and economic questions faster and on a much larger scale than any human journalist. Thus, automation could improve transparency and point journalists to important issues that need further attention. On the other hand, algorithms for generating automated news rely on data and predefined rules, which are subject to biases and errors. Publishers thus have to assure a certain level of algorithmic transparency and accountability (Diakopoulos, 2015), in particular, if the technology is used for critical and controversial topics.

Acknowledgements

Declaration of Conflicting Interests

Funding

References

Anderson C (2013) Towards a sociology of computational and algorithmic journalism. *New Media & Society* 15(7): 1005–1021. DOI: 10.1177/1461444812465137.

AP (2015) Apple tops Street 1q forecasts. *Yahoo Finance*, 27 January. Available at: Http://Finance. Yahoo.Com/News/Apple-Tops-Street-1q-Forecasts-213944804.Html

Carlson M (2015) The robotic reporter. *Digital Journalism* 3(3): 416–431. DOI: 10.1080/21670811.2014.976412.

Clerwall C (2014) Enter the robot journalist. *Journalism Practice* 8(5): 519–531. DOI: 10.1080/17512786.2014.883116.

Cohen S, Hamilton JT and Turner F (2011) Computational journalism. *Communications of the ACM* 54(10): 66–71. DOI: 10.1145/2001269.2001288.

Diakopoulos N (2015) Algorithmic accountability. *Digital Journalism* 3(3): 398–415. DOI: 10.1080/21670811.2014.976411.

Flanagin AJ and Metzger MJ (2008) Digital media and youth: Unparalleled opportunity and unprecedented responsibility. In: Metzger MJ and Flanagin AJ (eds) *Digital Media, Youth, and Credibility*. Cambridge, MA: MIT Press, pp. 5–27.

Golan GJ (2010) New perspectives on media credibility research. *American Behavioral Scientist* 54(3): 3–7. DOI: 10.1177/0002764210376307.

Gonçalo Oliveira H and Cardoso A (2015) Poetry generation with PoeTryMe. In: Besold TR, Schorlemmer M and Smaill A (eds) *Computational Creativity Research: Towards Creative Machines*, vol. 7. Atlantis Press, pp. 243–266.

Graefe A (2016) *Guide to Automated Journalism*. New York, NY: Tow Center for Digital Journalism, Columbia Journalism School.

Graefe A, Haim M, Haarmann B, et al. (2016) Replication Data for: Readers' perception of computer-generated news: Credibility, expertise, and readability. *Harvard Dataverse* V1. DOI: 10.7910/DVN/WHUEZA.

Haarmann B and Sikorski L (2015) Natural language news generation from big data. *International Journal of Computer, Electrical, Automation, Control and Information Engineering* 9(6): 1454–1460.

Hovland CI, Janis IL and Kelley HH (1953) *Communication and Persuasion: Psychological Studies of Opinion Change*. New Haven, CT: Yale University Press.

Johnson TJ and Kaye BK (2004) Wag the blog: How reliance on traditional media and the Internet influence credibility perceptions of weblogs among blog users. *Journalism & Mass Communication Quarterly* 81(3): 622–642. DOI: 10.1177/107769900408100310.

Kohring M and Matthes J (2007) Trust in news media: Development and validation of a multidimensional scale. *Communication Research* 34(2): 231–252. DOI: 10.1177/0093650206298071.

Lazer D, Kennedy R, King G, et al. (2014) The parable of Google Flu: Traps in big data analysis. *Science* 343(6176): 1203–1205. DOI: 10.1126/Science.1248506.

Leiner D (2014) *Convenience samples from online respondent pools: A case study of the SoSci Panel*. (Working paper). Germany: Ludwig-Maximilians-Universität Munich. Available at: https://www.researchgate.net/publication/259669050_Convenience_Samples_from_Online_Respondent_Pools_A_case_study_of_the_SoSci_Panel

Lemelshtrich Latar N (2015) The robot journalist in the age of social physics: The end of human journalism? In: Einav G (ed.) *The New World of Transitioned Media*. Heidelberg: Springer International Publishing, pp. 65–80.

Levy S (2012) Can an algorithm write a better news story than a human reporter? *Wired*. Available at: Http://Www.Wired.Com/2012/04/Can-An-Algorithm-Write-A-Better-News-Story-Than-A-Human-Reporter/

Lewis SC and Westlund O (2015) Actors, actants, audiences, and activities in cross-media news work. *Digital Journalism* 3(1): 19–37. DOI: 10.1080/21670811.2014.927986.

Metzger MJ, Flanagin AJ and Medders RB (2010) Social and heuristic approaches to credibility evaluation online. *Journal of Communication* 60(3): 413–439. DOI: 10.1111/J.1460-2466.2010.01488.X.

Meyer P (1988) Defining and measuring credibility of newspapers: Developing an index. *Journalism & Mass Communication Quarterly* 65(3): 567–574. DOI: 10.1177/107769908806500301.

Napoli PM (2014) Automated media: An institutional theory perspective on algorithmic media production and consumption. *Communication Theory* 24(3): 340–360. DOI: 10.1111/Comt.12039.

Pariser E (2011) *The Filter Bubble: What the Internet Is Hiding from You*. New York: Penguin Press.

Petrovic S and Matthews D (2013) *Unsupervised joke generation from big data*. Paper presented at the 51st annual meeting of the Association of Computer Linguistics, Sofia, 4–9 August.

Sundar SS (1999) Exploring receivers' criteria for perception of print and online news. *Journalism & Mass Communication Quarterly* 76(2): 373–386. DOI: 10.1177/107769909907600213.

Ulanoff L (2014) Need to write 5 million stories a week? Robot reporters to the rescue. *Mashable*. Available at: Http://Mashable.Com/2014/07/01/Robot-Reporters-Add-Data-To-The-Five-Ws/

Urban J and Schweiger W (2014) News quality from the recipients' perspective. *Journalism Studies* 15(6): 821–840. DOI: 10.1080/1461670x.2013.856670.

van Dalen A (2012) The algorithms behind the headlines. *Journalism Practice* 6(5–6): 648–658. DOI: 10.1080/17512786.2012.667268.

van der Kaa H and Krahmer E (2014) *Journalist versus news consumer: The perceived credibility of machine written news*. Paper presented at the Computation + Journalism Symposium, Columbia University, New York, 24–25 October.

West MD (1994) Validating a scale for the measurement of credibility: A covariance structure modeling approach. *Journalism Quarterly* 71(1): 159–168.

Westerman D, Spence PR and van der Heide B (2014) Social media as information source: Recency of updates and credibility of information. *Journal of Computer-Mediated Communication* 19(2): 171–183. DOI: 10.1111/Jcc4.12041.

Young ML and Hermida A (2015) From Mr. and Mrs. outlier to central tendencies. *Digital Journalism* 3(3): 381–397. DOI: 10.1080/21670811.2014.976409.

Author biographies

Andreas Graefe, PhD, is currently a research fellow at LMU Munich, Germany, and at Columbia University's Tow Center for Digital Journalism. He also holds the Professorship in Customer Relationship Management at Macromedia University. Andreas' research is quantitative-empirical,

interdisciplinary, and problem-oriented. He recently published the Guide to Automated Journalism and started to generate computer-written texts for the PollyVote.com project for forecasting US presidential elections.

Mario Haim, MA, is a research assistant at the Department of Communication Science and Media Research, Ludwig-Maximilians-University Munich, Germany. In his research, he focuses on all aspects of algorithmic influences on journalism. This includes data-driven journalism as well as personalization and Filter-Bubble effects, audience metrics, and news distribution within social network sites.

Bastian Haarmann works at the Fraunhofer Institute for Communication, Information Processing and Ergonomics, Germany.

Hans-Bernd Brosius (*1957) studied psychology and medicine at the Westfälische-Wilhelms-Universtaet in Münster. He received his doctoral degree there in 1983. From 1983 to 1995 he was a postdoctoral fellow and – later on – assistant professor for communication studies at the University of Mainz. He received his secondary doctoral degree (Habilitation) in 1994. Since 1996, he is professor of communication at the Institut für Kommunikationswissenschaft und Medienforschung, Ludwig-Maximilians-Universität München. From 1998 till 2002 he was chairman of the German Communication Association. Since 2001, he is also dean of the faculty of social sciences He is member of the editorial board a several leading communication journals. His research interests include media use, media effects, digitalization of mass media, and methodology.

Appendix 1

Main effects' analyses of variance.

Independent variable	Dependent variable	F	Pillai's V	Effect size r
Actual source		$F(3, 1848) = 119.2^{***}$	.16	
	Credibility	$F(1, 1850) = 75.8^{***}$		.20
	Readability	$F(1, 1850) = 146.1^{***}$		.27
	Expertise	$F(1, 1850) = 19.4^{***}$		.10
Declared source		$F(3, 1848) = 10.8^{***}$	.02	
	Credibility	$F(1, 1850) = 6.5^{*}$		.05
	Readability	$F(1, 1850) = 25.7^{***}$		.11
	Expertise	$F(1, 1850) = 1.9$		–
Real × declared		$F(3, 1848) = 0.9$	–	
	Credibility	$F(1, 1850) = 1.0$		–
	Readability	$F(1, 1850) = 0.8$		–
	Expertise	$F(1, 1850) = 0.1$		–

Results from one multivariate (overall) and three univariate (one per independent variable) analyses of variance (AnOVa). Rows containing independent variables and Pillai's V show multivariate results; rows containing dependent variables show univariate analyses. Results depict three levels of statistical significance (<.05*; <.01**; <.001***) and include all ratings (i.e. both topics; $n = 1854$).

Appendix 2

Results per topic.

		N	Credibility	Readability	Expertise
Finance		921			
Human written	Declared as journalist	232	3.3 (0.06)	2.8 (0.06)	3.0 (0.06)
	Declared as algorithm	228	3.1 (0.06)	2.6 (0.06)	3.1 (0.06)
Computer generated	Declared as journalist	231	3.7 (0.05)	2.5 (0.06)	3.2 (0.05)
	Declared as algorithm	230	3.7 (0.06)	2.3 (0.06)	3.1 (0.06)
Soccer		933			
Human-written	Declared as journalist	220	3.8 (0.05)	3.0 (0.07)	3.4 (0.05)
	Declared as algorithm	235	3.6 (0.05)	2.8 (0.06)	3.2 (0.05)
Computer generated	Declared as journalist	239	3.9 (0.05)	2.4 (0.06)	3.6 (0.05)
	Declared as algorithm	239	3.9 (0.05)	2.1 (0.05)	3.6 (0.05)

Mean ratings based on 5-point scales (1 = 'I completely disagree', 5 = 'I completely agree'), standard errors in parentheses.