Audience Evaluations of News Videos Made with Various Levels of Automation: A Population-based Survey Experiment

Neil Thurman,[i] *Sally Stares,*[ii] *and Michael Koliska*[iii]

The use of automation in news content creation is expanding from the written to the audio-visual medium with news organizations including Reuters turning to video automation services provided by companies such as Wibbitz. Although researchers have explored audience perceptions of text-based news automation, to date no published study has examined how news consumers perceive automated news videos. We conducted a between-subjects online survey experiment to compare how a socio-demographically representative sample (n=4,200) of online news consumers in the UK perceived human-made, partly automated, and highly automated short-form online news videos (n=42) on 14 different story topics. Our findings show that human-made videos received on average more favourable responses on some evaluation variables, although the differences were not large. We also found that there can be significant differences in the relative evaluation of automated and human-made news videos across different individual stories. For practitioners our results suggest partially automated news videos with post-automation human editing can be well received. For researchers our results show the need to use reasonably large sets of experimental stimuli, and suggest that maintaining socio-demographic variation within samples of respondents is worthwhile.

Keywords: AI, audience, automated journalism, news videos, perception, reception study, survey experiment, video automation

[i] Dept. of Media and Communication, LMU Munich <neil.thurman@ifkw.lmu.de>

[ii] Dept. of Sociology and Criminology, City, University of London <sally.stares@city.ac.uk>

[iii] Communication, Culture, and Technology master's program, Georgetown University <michael.koliska@georgetown.edu>

Introduction

The use of automation in journalism is encroaching more and more on what many would consider to be journalists' core professional practices, such as the identification of story leads, verification, and decisions about which stories are shown, and with what prominence (Thurman 2018; Fletcher, Schifferes, and Thurman 2020; Kunert and Thurman 2019). Over the past few years, automation has increasingly been used for the creation of written news texts too, and more recently the production of news videos has also become more automated (Fanta 2017). News organizations including the BBC, Reuters, and *The Economist* have turned to video automation services provided by companies such as Wibbitz, Wochit, and Synthesia.

This rise of automation in journalism has been linked to changes in audience demand for content (Dörr 2016), and the consumption of online news videos has been growing rapidly. While about 24% of news consumers across 40 countries watched online news videos on a weekly basis in 2016, this proportion had increased to 67% in 2020 (Newman et al. 2016; Newman et al. 2020). Video automation has helped to meet this growing demand, and significant time and resources are being invested in the development of applications that introduce elements of automation into the production of audio-visual content, including news. For example, Wibbitz—whose clients include Reuters, Condé Nast, *USA Today*, TMZ, and NBC—has raised over US$40 million, including from the Associated Press (Crunchbase n.d.-1), to develop products that, it says, allow the production of video "at unprecedented scale with the power of automation" (Wibbitz n.d.). And Wochit—whose news clients include Gannett, *La Presse*, Schibsted, *La Repubblica*, and the *Toronto Star*—has raised US$28.8 million, including from the parent company of PA Media, the UK's leading multimedia news

agency (Crunchbase n.d.-2), allowing it to offer "hands-free, data-driven, mass video creation" (Wochit n.d.).

An important aspect of understanding the impact of these new algorithmic tools within journalism is how audiences perceive news produced using automation. Although some progress has been made in exploring audience perceptions of text-based news automation (see Graefe and Bohlken 2020 for an overview), to date no known study has examined how news consumers perceive automated audio-visual news. Furthermore, many of the existing studies on the perception of automated journalism have been limited by small and unrepresentative samples of respondents, questionable measures of the dependent variables, and a quantity and quality of experimental stimuli that has made the isolation of authorship (automated or human) as an independent variable problematic. These limitations make it difficult to know how news consumers really evaluate automated news in comparison with human-made equivalents, and to what extent those judgements might be generalizable.

This study attempts to help fill this twin research gap in two ways. Firstly, we design and conduct a between-subjects experiment to compare how online news consumers in the UK perceive human-made, partly automated, and highly automated short-form online news videos. Secondly, we employ a more rigorous methodology than has been seen in the literature to date: we undertake a qualitative study to develop a comprehensive range of dependent variables; to address the fact that subject matter may affect results we use 14 sets of videos, each on a different story topic, and each containing a human-made, partly automated, and highly automated version; and we use a large sample of respondents (n=4,200), recruited by a market-leading polling company, that is representative of the UK adult population of online news consumers[1]

in terms of key socio-demographic variables. In these ways we deliver study results with unprecedentedly high internal and external validity.

Overall, the findings of this study show, firstly, that human-made videos have the edge over those created with automation, with human-made videos receiving more favourable responses on a range of evaluation variables. However, the differences we found are not large, and in some cases there were no significant differences, leading to our conclusion that a modest use of automation, with some post-automation human editing, can be well received. Secondly, our results show that there can be significant differences in the relative evaluation of automated and human-made news videos across different individual stories (e.g. Winter Olympics vs cryptocurrency vs Turkish election), suggesting that researchers ought to avoid the practice of using small sets of experimental stimuli. Thirdly, we found limited evidence of socio-demographic variables interacting with relative evaluations of human and automated journalism. Future research would benefit from maintaining socio-demographic variation with samples in order to add to the evidence base on this question, which has received little attention to date.

Literature Review

So-called "automated journalism" focuses on the production of news content and may be seen as a sub-category of computational journalism, by which is meant "the advanced application of computing, algorithms, and automation to the gathering, evaluation, composition, presentation, and distribution of news" (Thurman 2019).

Carlson's (2015) definition of automated journalism as "algorithmic processes that convert data into narrative news texts with limited to no human intervention beyond the initial programming" (146) has been widely adopted but is no longer strictly

accurate. Firstly, automation, as discussed, is now being used to create audio-visual news as well as textual news. Secondly, journalists have begun to post-edit automated outputs prior to publication. Although this post-editing is not currently, or ever likely to be, universal, it is happening with increasing frequency. For example, the data-driven stories produced, using automation, by the RADAR News agency and sent to their local and regional newspaper clients in the UK are routinely post-edited by journalists at those newspapers before being published (Renwick 2020). Thirdly, automated journalism is now working with inputs—like moving images—other than the textual and numeric "data" Carlson's (2015) definition mentions. For example, there now exist forms of automated video journalism that, taking raw video footage as input, use artificial intelligence to create highlight videos suitable for broadcast (Wimbledon n.d.). Other video automation technologies, such as those offered by Synthesia and Wibbitz, can turn text into news videos.

For these reasons, and building on Carlson (2015), we define automated journalism as: "Algorithmic processes that convert numerical data, images, or text into written or audio-visual news items with various levels of human intervention beyond the initial programming."

Research into the perception of automated journalism has, at the time of writing in March 2023 and as far as we are aware, focused solely on written news texts. Broadly, these studies have one or both of two main aims. Firstly, to compare the perceptions of news texts actually or purportedly generated by humans or machines, and, secondly, to explore factors, mainly psychological, that may explain any variance in these perceptions. These factors have included audience's expectations of journalists and automated systems (Waddell 2018); the credibility of the news brand carrying the texts (Liu and Wei 2019); the level of transparency given about how the news item was

automated (Graefe, Haim, and Diakopoulos 2017); and the cultural (Zheng, Zhong, and Yang 2018), professional (Van der Kaa and Krahmer 2014; Jung et al. 2017), or other demographic (Melin et al. 2018) characteristics of the respondents. Surprisingly, perhaps, there appear to have been few attempts to explore the extent to which the technical characteristics of articles—such as their narrative structure, inclusion of quotes, or length—which may well differ between automated and manually produced articles, explain any differences in evaluation. Tandoc, Yao, and Wu (2020) provide a partial exception, as they examined whether the perception of articles—declared as written by either a machine, a human, or both—that were "objective" (containing direct, sourced quotes, and neutral language) or "non-objective" (containing paraphrased quotes and "value-laden" words) differed, although the "automated" journalism used in the study was only declared as such, and was actually written "by a former journalist" (554).

A meta-analysis (Graefe and Bohlken 2020) of 12 of the studies published up to October 2019 found that there were no differences in readers' perceptions of "credibility" between the human- and machine-written news articles. However, human-written news was perceived as being of slightly higher "quality" and much more "readable". Together, the studies also found that people rated articles more highly across these three criteria if they were told the article was written by a human, even if it was not.

This analysis might suggest that, in our study, we may find some differences in the perception of news videos made manually and by machine, but that those differences may not be great and may even be non-existent by some criteria. This is, perhaps, a simplistic expectation given that the automated production of textual and video news differs a great deal.

Machine-written news texts rely primarily on natural language generation (NLG), translating structured data into text using pre-written templates or linguistically sophisticated rule-set systems (Diakopoulos 2019). These templates can be fixed, but

> in order to add variability to the final outputs, individual words in the template can be assigned an unlimited number of synonyms that are picked at random when the texts are generated […] and branching offers the ability for words, sentences, or whole paragraphs to be inserted depending on certain conditions. (Thurman, Dörr, and Kunert 2017, 1245)

Natural language generation systems are thus able to manipulate both the semantics (content) and syntax (arrangement) of a written text, according to the linguistic norms and standards of a specific language.

In contrast to written texts, the visual language of videos does not follow similarly precise linguistic or widely agreed upon syntactic rules. While videographers and filmmakers follow certain conventions in visual storytelling (such as the use of wide shots that provide a sense of location or close-ups to provide more intimate depictions of the objects and subjects on screen), these conventions can be relatively freely employed. The lack of strict visual rules means that the algorithmic production of videos takes a different approach to that used for text. For one variant of algorithmic video production—text-to-video automation—the narrative structure of words (spoken or used as captions) is the starting point, with the algorithm matching the words with the metadata that classify images within a data bank.

The fundamental differences between news texts and videos also mean that the criteria by which they are judged will differ. Some of the criteria used to compare human- and machine-written news texts, like “pleasant to read” and “well-written” (see, e.g., Clerwall 2014), simply do not apply, while others that have not been used, such as how well images match the captions or the use of audio, may. Another reason why the

results of this study may differ from the findings of prior studies on the perception of automated text journalism is that the quality of the experimental stimuli used in some of those studies has made the isolation of authorship (automated or human) as an independent variable difficult. For example, Clerwall's (2014) study compared the evaluations of an automated factual report on an American football game against the evaluations of a human-written opinion column about the prospects of three quarterbacks. The differences he found between the evaluations of the automated and human-written stories could, therefore, have been partially, or even fully, a consequence of their genre (match report vs opinion) and not wholly or even partially a consequence of their authorship (automated vs human).

As Sundar (1999) has written, "people may not always carry values in their heads for all the measures that interest researchers" (374). This may have been the case in prior studies on the perception of automated journalism, which have often adopted measures developed without reference to automated journalism and even without reference to journalism of any form. For example, van der Lee et al.'s (2018) study on the perception of automated sports journalism exclusively used measures developed by Maes, Ummelen, and Hoeken (1996) for the evaluation of "instructive texts".

Because our object of study and methodological approach differ from previous perception studies of automated journalism, we considered that developing a set of hypotheses from previous literature, using measures that might not be relevant, and setting any expectations about effect sizes, was premature, and that we needed to adopt a more open, inductive, exploratory approach for this study. Fundamentally this still indicated an experimental design, in the sense of randomly allocating participants to watch and evaluate differently authored videos, but not in a traditional hypothetico-deductive framing. Instead of posing hypotheses loosely derived from a slightly

different field of enquiry, we decided not to develop and test hypotheses on the relative evaluations of human-made and automated news videos but rather to ask this general research question:

> **RQ1**: What, if any, differences exist in UK online news consumers' evaluations of short-form online news videos made with various levels of automation, and none?

Many of the existing studies on the perception of automated journalism use small samples of stimuli. For example, the studies by Clerwall (2014) and Jung et al. (2017) each used one pair of stories (on sport), while Wölker and Powell (2021), Graefe et al. (2018), and Li et al. (2022) each used two pairs of stories. As a result, it is far from clear how much a story's topic will influence the relative evaluations of the human and automated versions. Therefore, our second research questions asks:

> **RQ2**: How, if at all, do any differences found in RQ1 vary across the 14 story topics (e.g. Winter Olympics vs cryptocurrency vs Turkish election) included in our experiment?

Prior studies on the perception of automated journalism have often used samples of college students that skew young and female. For example, the respondents in Clerwall's (2014) and Jia's (2020) studies had a mean age in the 20–29 range and were predominantly female. There are some indications, however, that socio-demographic characteristics can make a difference. For example, Melin et al. (2018) found that young women in their small sample (n=152) liked automated news articles significantly less than older men did. There is, however, a lack of robust evidence about the variability in how individuals of different socio-demographic groups evaluate automated news. Comparing how, if at all, our results from RQ1 differ across socio-demographic groups would not only increase the external validity of our findings but could also contribute, more generally, to the evidence about the extent to which, if at all, differences in

evaluations of automated and human-made news vary along socio-demographic lines. Logg, Minson, and Moore (2019; Logg 2017) found that algorithmic appreciation held true across different applications. Although not investigated by Logg, Minson, and Moore (2019), this may also be so of different news story formats, such as text and video. Therefore, our third research question asks:

> **RQ3**: How, if at all, do any differences found in RQ1 vary according to the gender, age, work status, region, and social grade of respondents?

Methodology

Experimental Design

A 14 (video topic) x 3 (level of automation) between-subjects population-based online survey experiment was conducted, with each participant viewing a single video (n=100 for each video). A between-subjects design was considered preferable to a within-subjects design (where participants would watch and evaluate several videos) to avoid results being potentially affected by learning or priming effects: we wanted respondents to give an independent evaluation of any video they watched rather than their responses to one video being influenced in some way by their experience of watching a previous one (Charness, Gneezy and Kuhn, 2012). Given that the sample was drawn from survey panels that included wider ranges of individuals than just engaged college students,[2] we also wanted to set a very low threshold for respondent fatigue. Keeping the task very short was therefore important. Since this was an exploratory study, and given the lessons learned from Graefe and Bohlken (2020, described above), we had no prior information on what effect sizes should be considered noteworthy, which made it impossible to conduct a meaningful power calculation to specify our sample size. We arrived at n=100 per video largely pragmatically, choosing a larger sample size than

typically used in relevant literature described above. The study was approved by the Ethics Committee of the Department of Sociology at City, University of London. As Mutz (2011) states, population-based survey experiments combine the best aspects of large-scale observational surveys and laboratory-based experiments while "eliminating many of their weaknesses". In particular, they allow a determination of causal effects and the generalizing of any resulting conclusions to a wider population.

Stimulus Materials

Human-made Videos

The human-made videos were sourced from PA Media (PA), the national news agency for the UK and Ireland. The PA "consumer-ready" videos used in this study were captioned and had been created without the help of automation (Alex Rothwell, PA's Head of Video, personal communication).

Videos Produced Using Automation

The partly and highly automated videos were created using the Wibbitz platform. At the time the videos were produced, the Wibbitz production process started with text. Users of the platform could decide whether the text should appear as captions or be used as a script for a voice-over. The platform automatically tried to find media—both video clips and still images—that matched the captions or voice-over by searching media databases, including those provided by Getty, Reuters, Pond5, and WENN. By default, background music was added. Each video used one of a variety of "themes", pre-built stylistic templates that varied in their use of transitions between scenes (cuts, wipes, dissolves, etc.); colour overlays; the colour, size, and style of the captions (if used); and the size and aspect ratio of the video itself.

Although the Wibbitz platform could produce videos of an acceptable quality "out of the box", with no human intervention beyond the initial text input, in our experience most videos needed some further editing before they were consumer-ready. The platform provided tools that allowed the initial, autonomously produced videos to be edited manually. Editors could add or subtract complete scenes; insert and delete still and moving images; change the size, alignment, colour, and style of textual captions, as well as the text itself; change the soundtrack; include or exclude a company logo; and change the colour palette.

Because the aim of this research project was to compare UK citizens' evaluations of short-form news videos 1) made by journalists and 2) made using automation, a key issue was the degree of automation used. As should be clear from our description, the videos produced on the Wibbitz platform could range from the highly automated (relying only on an initial text) through the partially automated (those that have undergone some further human editing) to the minimally automated (those that have undergone heavy further human editing). Typically, the Wibbitz platform was used to partially, rather than highly, automate video production. Hilary Kay, then VP of Marketing at Wibbitz, said that for the content they were strongest in—international news, politics, finance, sports, entertainment, and technology—the videos produced by the platform out of the box were "very high quality about 85 to 90% of the time". For content that the platform was "not as strong in, for example lifestyle", the videos were "between 60 and 80% of the way there" (personal communication). In those cases, "editors would come in and make changes" to make them ready for publication. The typical changes made involved deleting duplicate images or inserting an illustrative video that the algorithm did not find.

In order to reflect the different degrees of automation with which videos on the Wibbitz platform can be produced, we created both highly and partially automated videos. To try to ensure the Wibbitz platform was not handicapped in terms of its ability to find suitable illustrative images in comparison to the video journalists at PA, we did not select PA videos if they were about local or regional UK topics or topics that were otherwise so UK-focused that it would have been unlikely that the US-based Wibbitz would have had access to relevant still images or videos.

We identified potentially suitable videos from PA then transcribed their captions and inputted the resulting texts into Wibbitz. The platform then created automated videos based on these textual inputs. With this method, the human-made and automated videos shared the same underlying text input, ensuring that the textual content of the captions was not an independent variable. A disadvantage of this method is that it could put the human-made PA videos in an advantageous position. The captions written by PA journalists would have been determined, in part, by the illustrative video clips and/or still images they had available. The Wibbitz platform does not have access to the same set of images as PA. This is one possible explanation for why the Wibbitz platform was unable to find enough—or in some cases any—video clips and/or still images to illustrate some of the videos it produced from the transcribed PA caption texts. To level the playing field, we did not use highly automated Wibbitz videos produced by this method if they contained many empty frames or if not a single still image or video clip in the Wibbitz video matched the story topic.

When the Wibbitz platform automatically produced a video with still images or video clips in most frames, and with at least one of those images matching the topic of the text input, then that video was shortlisted for inclusion in our experiment as a "highly automated" video. However, because, as we have stated, the initial videos

produced by the Wibbitz platform typically undergo further human editing, we also created a set of "partly automated" videos from the "highly automated" videos. When converting the highly automated videos to the partly automated videos, we manually replaced still images and video clips that did not match the captions. Across the 14 partially automated videos used in our experiment, an average of 55% of the scenes were edited (SD=12.5) (see Table A in the supplemental material).

Because the independent variable in our survey experiment is the degree of automation used to create short-form news videos, it was important to try to minimize the extent to which other variables could have an effect, with these variables including the video topic and stylistic elements such as music, editing transitions, and captions. In order to do this we used 14 sets of videos covering topics—royalty, sport, celebrity, business, politics, crime, technology, and culture—familiar to UK audiences, specifically:

1. Pippa Middleton, sister of the Duchess of Cambridge, announces she is expecting a baby.
2. A review of Cristiano Ronaldo's career in 60 seconds.
3. Elon Musk apologizes for making accusations against British cave diver.
4. Recep Tayyip Erdogan wins Turkey's presidential election.
5. Harvey Weinstein pleads not guilty to rape in New York court hearing.
6. Canadian singer Justin Bieber announces his engagement to Hailey Baldwin.
7. Prince Louis of Cambridge is christened.
8. Donald Trump's commemorative star on the Hollywood Walk of Fame is destroyed.
9. Milan–Cortina is chosen to host the 2026 Winter Olympic and Paralympic Games.
10. An explainer video on 5G wireless mobile technology.
11. An explainer video on Facebook's cryptocurrency, Libra.

12. Northern Irish play The Ferryman wins best play at the Tony Awards.
13. Theresa May announces her decision to step down as Conservative Party leader and UK Prime Minister.
14. A video highlighting numerical facts about the Wimbledon tennis championships.

Each set containing a human-made, a partly automated, and a highly automated video on the same story (see Table A in the supplemental material). Furthermore, we made sure that both the human-made and automated videos did not differ in their resolution and contained nothing (e.g. logos) that would give away their authorship.

In addition, we made sure that the human-made and automated videos did not differ significantly stylistically. The PA videos were fairly consistent in style, using white captions and simple transitions between scenes (mostly cuts) and with no colour overlays or narration. In order that the Wibbitz videos matched this style, we created only captioned videos on the Wibbitz platform, not voice-over videos; set the editing style ("Theme Type") to "Minimal" to ensure simple transitions between scenes with no colour overlays; made all the captions white, with no colour highlights; ensured the size and position of the captions were broadly similar to those used by the PA videos; and ensured background music of a similar style to that used by the PA videos was used.

Not all of the PA videos used music: six of the 14 used background sound instead. The equivalent partly and highly automated videos used music, as no suitable background sound was available. The length of the videos averaged 63 seconds (SD=21.8). The human-made videos were, on average, slightly shorter (M=56, SD=16) than the partly and highly automated videos (M=66, SD=24). The human-made videos relied more on moving images, with eight of the 14 PA videos featuring all moving images and six a mixture of still and moving images. By contrast, 57% of the highly and partly automated videos contained just stills and the other 43% a mixture of stills and

moving images (see Table A in the supplemental material). All the videos used in the experiment are available at (Thurman et al. 2023).

Survey Development

Mindful of Sundar's (1999) warning about the validity of the dependent variables used in prior news perception studies, we undertook a major project with an inductive (rather than theory-driven) approach to better understand the criteria with which online news consumers evaluate online news videos. In-depth group interviews were held with a socio-demographically diverse sample of 22 online news video consumers in the UK. Each group was shown several online news videos, both human-made and made with the help of Wibbitz's automation platform. Subsequently, interviewers facilitated discussions about participants' reactions to and perceptions of the videos, which were recorded and transcribed verbatim. The resulting 145,000 words were analysed using thematic analysis. The results (Koliska et al. 2021) show the complex and interwoven set of criteria used to evaluate online news videos and were used to develop this study's survey instrument. The correspondence between the items we designed from our qualitative study, and the evaluative criteria used in the literature outlined above, is something of a Venn diagram. For example, previous studies have gauged audience evaluations of "readability" and "clarity", while one of the criteria that emerged from our group interviews was narrative flow. One respondent talked about videos "having flow" and another liked videos that had "a beginning, a middle and an ending". So, the final survey included two questions on narrative flow using semantic differential scales, where respondents were asked to place a slider towards the statement that best described their opinion of the video. One item ranged from "the story had a structure" to "there was no structure to the story", and another ranged from "the story flowed" to "the story

was disjointed". We also used this item format to ask about respondents' evaluations of bias/fairness, subjectivity/objectivity, and accuracy/inaccuracy, which correspond to the broad theme of "credibility" used in previous studies.

A close study of human-made and partly and highly automated videos also informed the development of the survey. For example, we noticed differences in how well illustrative images matched captions; the proportions of still and moving images used; the proportions of speech and (non-musical) background audio used; the range, quality, and specificity of the illustrative imagery used; the use of background music; and the use of captions. Therefore, questions were included in the final survey on such themes, which had also been discussed in the group interviews. For example, one question asked "How well or badly did the still images and/or moving footage match the words in the captions?" The response scale for this question—as was the case, where possible, with all the questions in our survey—explicitly displayed the evaluative dimension (e.g. "They all related well to the story") to avoid the reliability and validity problems, now extensively documented (see, e.g., Krosnick and Presser 2010, 275–278), with defaulting to the application of Likert-type agree-disagree response scales for all items. The items used as our dependent variables are given in Tables 1, 2, 3, and 4; and the full survey, including question wording and logic, is included in the supplemental material.

Expert-driven Pretesting

Presser and Blair (1994) found that "expert panels were the most productive method" in identifying problems in surveys. The first draft of our survey received expert review from a number of professional survey experts at a leading market research organization, Ipsos MORI, resulting in changes in the wording and logic of the survey.

Table 1: Survey items relating to how much respondents disliked or liked videos (continuous scale)

How much did you like or dislike…	−3 Disliked very much	0 Neither liked nor disliked	3 Liked very much
...the video?			
...the use of captions in general?			
...the still images and/or moving footage that was shown within the video?			
...the video's audio track?			
...the music in the video? [only asked for videos with music]			

Table 2: Survey items using a semantic differential format (7-point integer scales)

For each pair of statements, please move the slider towards the statement that best describes your opinion of the video:		
	−3	+3
inaccurate		accurate
biased		fair
fact-filled		opinion-based
subjective		objective
told me things I didn't know		told me things I knew already
presented only some of the relevant facts		presented all of the relevant facts
the story had a structure		there was no structure to the story
the story flowed		the story was disjointed
relatable		not relatable
informative		uninformative
professional		amateur
dry		emotive
serious		light
engaging		boring
understated		sensational
confusing		clear
comprehensive		limited

Table 3: Survey items relating to video captions and still/moving images

For the characteristics stated below, please select one response that best represents your personal opinion on the video overall	Responses					
Readability of captions	Difficult to follow overall	In some places difficult, in other places easy to follow	Easy to follow overall	Don't have an opinion		
Size of the text	Too small	About the right size	Too big	Don't have an opinion		
Amount of text	Not enough	About the right amount	Too much	Don't have an opinion		
Speed of movement/transition of captions	Too slow	About the right speed	Too fast	Don't have an opinion		
How relevant did you think the still images and/or clips of moving footage were to the story?	None related well to the story	Only some related well to the story	There was a roughly even mixture of related and unrelated ones	Most related well to the story	They all related well to the story	Don't know
How would you rate the quality of the still images and/or moving footage? By quality we mean things like how they were framed, their lighting and so on.	Very poor quality	Poor quality	Of mixed quality	Good quality	High quality	
How well or badly did the still images and/or moving footage match the words in the captions?	Not at all well	Not well	Satisfactorily	Well	Very well	Don't know
And what about the variety of still images and/or clips of moving footage within the video?	They were too similar to each other	There was a good balance between similar and varied ones	They were too varied			

Table 4: Survey items relating to audio characteristics of videos, and video authorship

Questions	Responses		
Would you have liked to have heard more of any of these elements in the video you watched?			
Location sound	No	Yes	
Music	No	Yes	
A voiceover from a journalist	No	Yes	
It was fine as it was	No	Yes	
Which of the following statements best matches your opinion on the use of music? The music...	Was suited to the news story	Was not suited to the news story	Don't know
The use of music…	Worsened my experience of the news story	Had no impact on my experience of the news story	Enhanced my experience of the news story
These days there are computer programmes that can automatically generate short news videos based on a text story provided by a journalist. How do you think the video you watched was made?	Completely made by a journalist (human made)	Mostly made automatically by computer software and then further edited by a human journalist (partly automated)	Mostly made automatically by computer software (highly automated)

Respondent-driven Cognitive Pretesting

Cognitive pretesting of the first draft of the survey was carried out with 10 participants. Respondents were recruited to ensure a mix of genders (40% were female), ages (M=45.5, SD=15.4), ethnicities, socio-economic backgrounds, and professions.

Two researchers conducted the pretests. Each participant spent an hour alone with a researcher in a university classroom. Respondents were asked to take the online survey, vocalizing their thoughts about the questions—and their answers—as they did so. The researchers asked probing questions at various points in order to further explore respondents' reactions to the survey. A number of the pretests were conducted using smartphones, and a number using computers. Because different videos are likely to prompt different reactions, none of the participants viewed the same video as part of the

survey they took, which meant that the appropriateness of the survey questions and response scales could be more fully explored.

The researchers made notes and the pretests were audio-recorded. Following analysis of the notes, a number of changes were made to the survey. These included refining some of the instructions given, rewording elements of some questions, and amending the logic and presentation of the survey.

Survey Scripting

The survey was scripted by Ipsos using their proprietary online survey platform. Respondents were informed that they had to be online with sufficient bandwidth to watch a video in order to take the survey. The survey included a question about how the video played. If a respondent reported that the video did not play smoothly they were not invited to complete the survey. Respondents were informed that they had to have the sound turned on when they watched the video and be able to listen to the sound (through headphones or speakers). The survey included a question in which respondents were played an easily identifiable sound. Those who could not identify it were assumed to be unable to listen to audio and were not invited to complete the survey. A multiple-choice attention check question was included at the end of the survey that asked about the content of the video the respondents had just watched. Respondents who failed to recall the subject of the video (e.g. "5G mobile technology" or "the Winter Olympics") were removed. To encourage respondents to watch the video stimulus fully, the survey could not be advanced until the video had finished playing. The stimulus was also included at the bottom of some of the question pages so respondents could, if they wished, watch it again when answering the survey questions. In order to counter response order effects (see, e.g., Krosnick and Presser 2010), two strategies were adopted, where possible. Firstly, for questions with response scales that did not have a

logical progressive ordering, the order of response options was randomized for each respondent, and, secondly, for questions with response scales that did have a logical progressive ordering, the response scale was reversed for alternate respondents—see questionnaire document in supplemental material for details.

Sample Recruitment and Composition

Respondents were recruited by Ipsos using a blend of its own panel of online survey respondents, and those of an approved partner provider. Participants in such panels are those who have signed up to take part in various online surveys for some compensation. They are not therefore a strict probability sample of the general population, but when the panels are large enough (commercial panels usually run to many thousands), they can provide considerable socio-demographic diversity. Our target population was in any case not the general public, but UK adults who consume online news. Quotas were employed so that in each cell of the experiment the sample provided a close match to the wider online UK population (those who use the Internet at least once a week) in terms of distributions of gender, age band (18–34, 35–54, 55–75), and region of residence (North, Midlands, and South) (see Table B in the supplemental material); a screening question was used to select into the sample only those who consume online news at least once a month.

Each of the 42 videos was watched by a different group of 100 respondents, giving a final sample size of 4,200. The 100 respondents in each group were randomly allocated to each condition using a least fill function on the three quota variables to reduce any skew on these characteristics occurring by chance, and it was ensured that fieldwork for each video topic was conducted over the same precise period in order to avoid introducing any bias from the time of day or day of week when participants

engaged with the survey (this was judged by Ipsos to be the main potential bias across cells).

Pilot Testing

A soft launch test with approximately 100 respondents was conducted prior to full fielding. No problems were found.

Data Quality Assurance

Ipsos carried out a variety of checks on the responses (including IP checks, bot checks, checks on "speedsters" and "straight liners", and checks to avoid multiple responses from single respondents), rejecting any that failed quality assurance.

Data Analysis

We analysed the data using SPSS software. For RQ1 and RQ2 we employed one-way ANOVAs and independent samples t-tests to compare mean scores for questions with a continuous answer scale between the experiment conditions, and chi-square tests to compare percentage responses to the categorical items between experiment conditions. For RQ2 we split the data by video topic and repeated the t-tests and chi-square tests within each of the 14 sets. This is tantamount to allowing, in each analysis, an interaction between video authorship, evaluation variable, and video topic (i.e. allowing the association between video authorship and evaluation variable to be different across different topics). For RQ3 we focused on the continuous items as dependent variables and modelled this more formally in a series of linear regressions, with video authorship, content, and socio-demographic items as explanatory variables, testing interaction terms between video authorship and each of the other explanatory variables. We analysed all items individually rather than combining them into indices because we wanted to retain

the granular detail of the separate characteristics studied. Exploratory factor analyses suggested some intuitively plausible underlying dimensions in the data, but with a few anomalies.[3]

Results

RQ1: What, if any, differences exist in UK online news consumers' evaluations of short-form online news videos made with various levels of automation, and none?

To summarize our results broadly: we found, on a number of evaluation items, that respondents gave more positive ratings, on average, to human-made than to automated videos. No differences were very large, however (the largest mean difference is 0.25 on a scale with range 7), and not all were statistically significant (at conventional levels). Figure 1 (and Table C in the supplemental material) summarizes these results across (pooling data from) all of the 14 sets of news videos. It shows, for example, that average scores for overall liking of the videos were 0.13 points lower, a statistically significant difference ($p<0.05$), for highly automated videos than for human-made videos, but that there was no statistically significant difference in overall liking scores between partly and highly automated videos, nor between partly automated and human-made videos. We found that liking of audio track, music (where used), and still images/moving footage was significantly lower for the partly and highly automated videos than for human-made videos. Reflecting how the automated videos in our experiment were created, the only significant difference between the highly and partly automated videos in the liking measures was the liking of images/moving footage. (The partly automated videos were versions of the highly automated videos that had been post-edited by a human to replace any images that did not match the captions.) Although liking of captions was significantly higher for human-made than for highly automated videos, it

was not so for either of the other two comparisons—this again making sense in the context of the experimental stimuli, in which the content of the captions was identical between experiment conditions.

For the items that asked about characteristics in terms of pairs of opposing evaluations (e.g. "engaging – boring", "professional – amateur"), we also found very few significantly different ratings between the partly and highly automated conditions. Our focal interest is the comparison between partly automated and human-made videos (given that, in practice, automated videos are rarely used without some element of human editing) (see Figure 1 and the first column of results in Table C in the supplemental material). Human-made videos tended to be rated more favourably in terms of telling respondents things they didn't know; their professionalism; fact-filledness; engaging nature; comprehensiveness; story flow; and understatedness—they were also on average rated as more emotive and lighter in tone. By contrast, there were no significant differences in terms of story structure; number of relevant facts presented; bias/fairness; subjectivity/objectivity; relatability; informativeness; clarity; or accuracy. Arguably, some of the items for which significant differences were found relate to some of the subtle qualities of human editing—such as professionalism, level of engagement, story flow, and understatedness—while several of the others—like number of relevant facts, level of bias, informativeness, and accuracy—would naturally be judged on the basis of the caption content (which did not vary by experimental condition). Relatability would likely be judged in relation to the video topic, which again was constant over experimental conditions.

In a similar vein (summarized in Figure 2), associations between experimental condition and judgements of readability of captions and amount of text were non-significant.

However, participants were significantly more likely to judge the size of the text to be too big or small in human-made videos than in automated videos, but also significantly more likely to judge that text transitioned at "about the right speed" (rather than too fast or slow). The images/moving footage were rated significantly more positively in human-made than automated videos in terms of relevance, matching to captions, and variety. When respondents were asked if any different audio effects would have been welcome, increased levels of automation returned significantly higher votes for (more) music and voice-over. Participants who watched a human-made video were most likely to rate the audio as "fine as it was". And finally, we found evidence that, at least at an aggregate level, differences in authorship were identifiable. Respondents were asked to guess whether the video they watched was human-made, partly automated, or highly automated, and although a majority (60%) guessed "partly automated" for each condition, the judgement of "human-made" was more likely for the human-made videos than for the automated ones, and the judgement of "highly automated" was more likely for the highly automated videos than for the human-made ones.

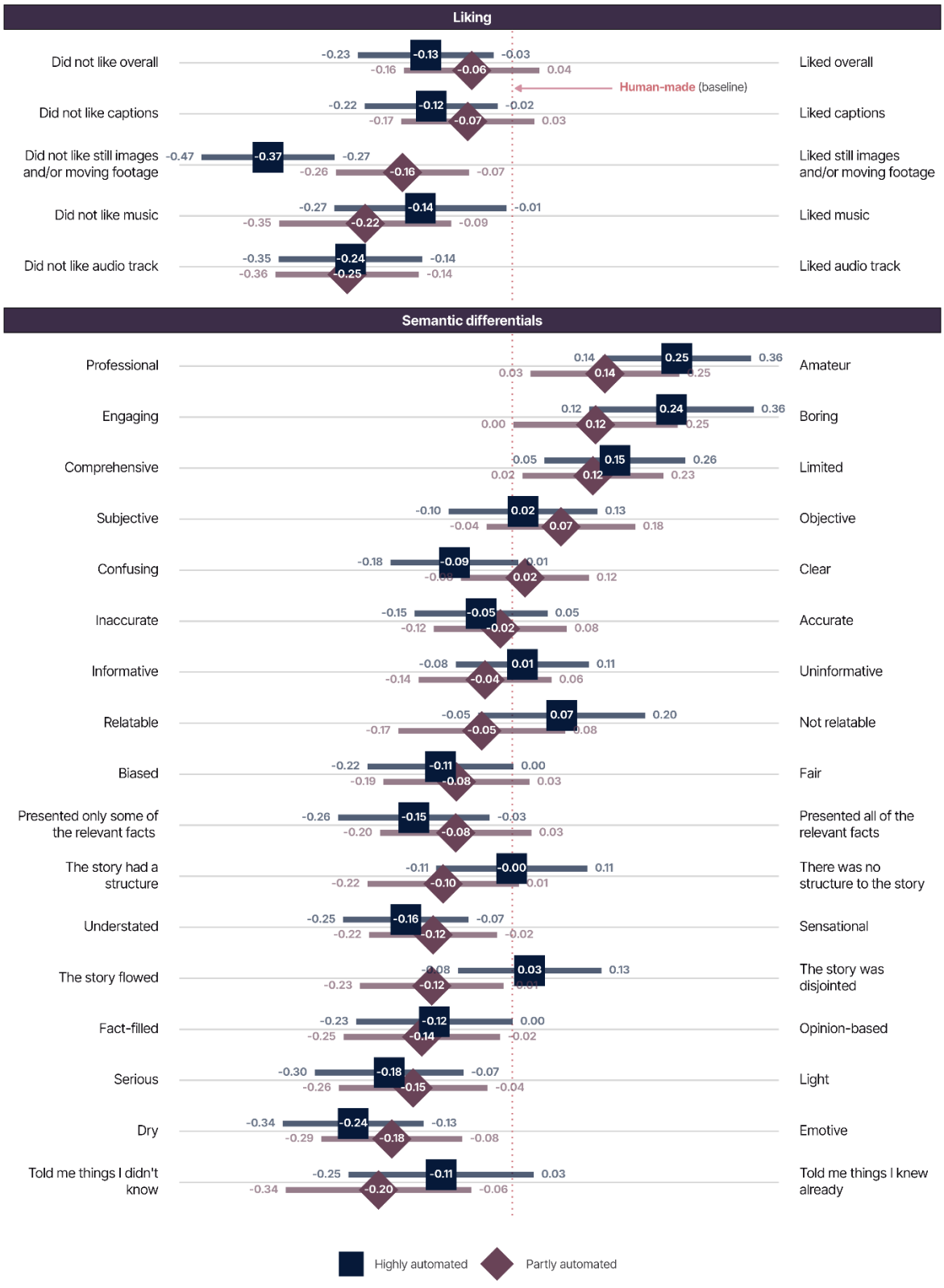

Figure 1: Differences in mean scores of liking and semantic differential items for highly automated (square) and partly automated (diamond) videos compared to human-made videos; 95% confidence interval in whiskers; scales each have a range of 7.

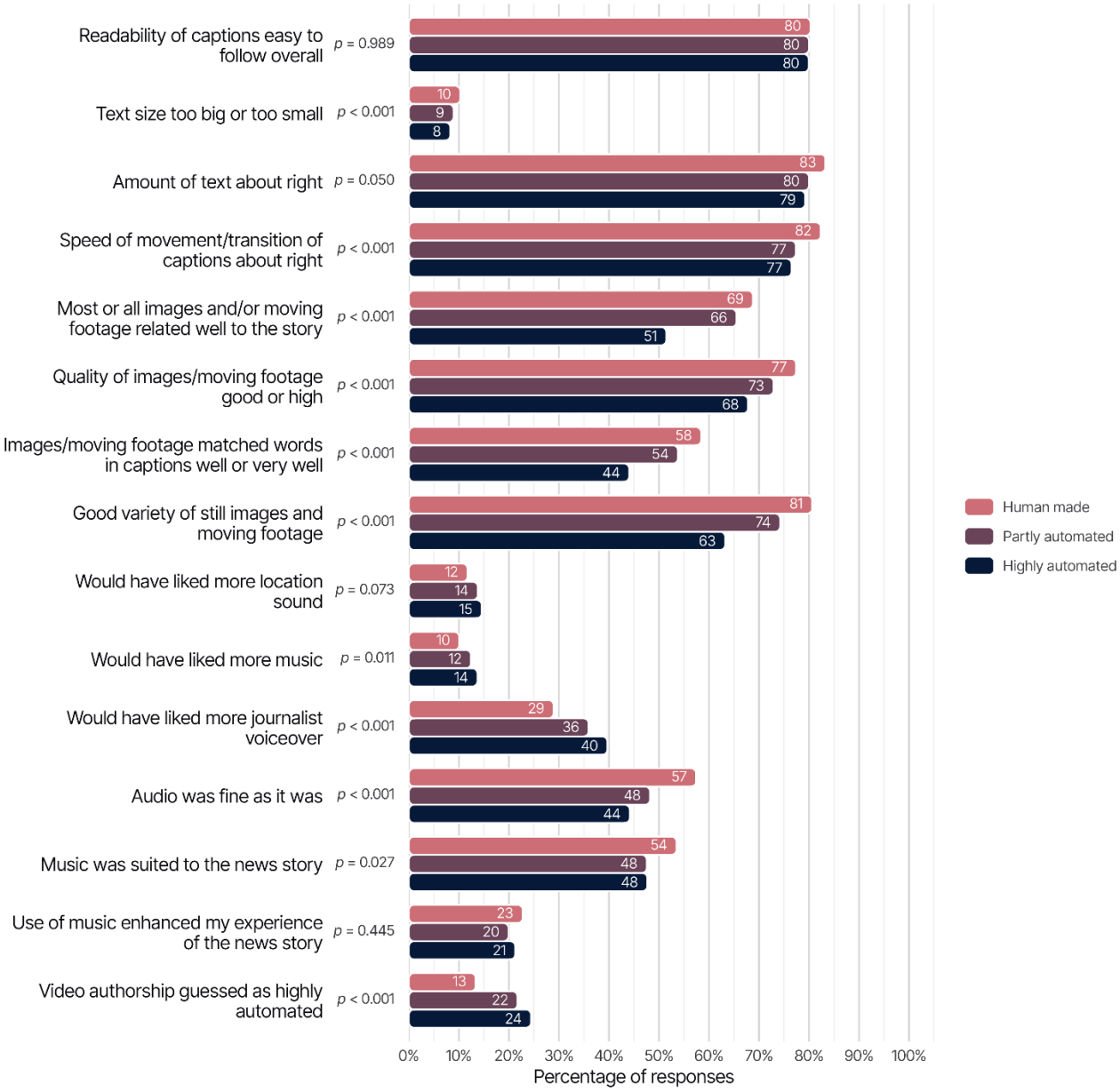

Figure 2: Percentages of answers to selected response options for the categorical survey items, with p-values from chi-squared tests of association between survey item (all categories) and level of automation.

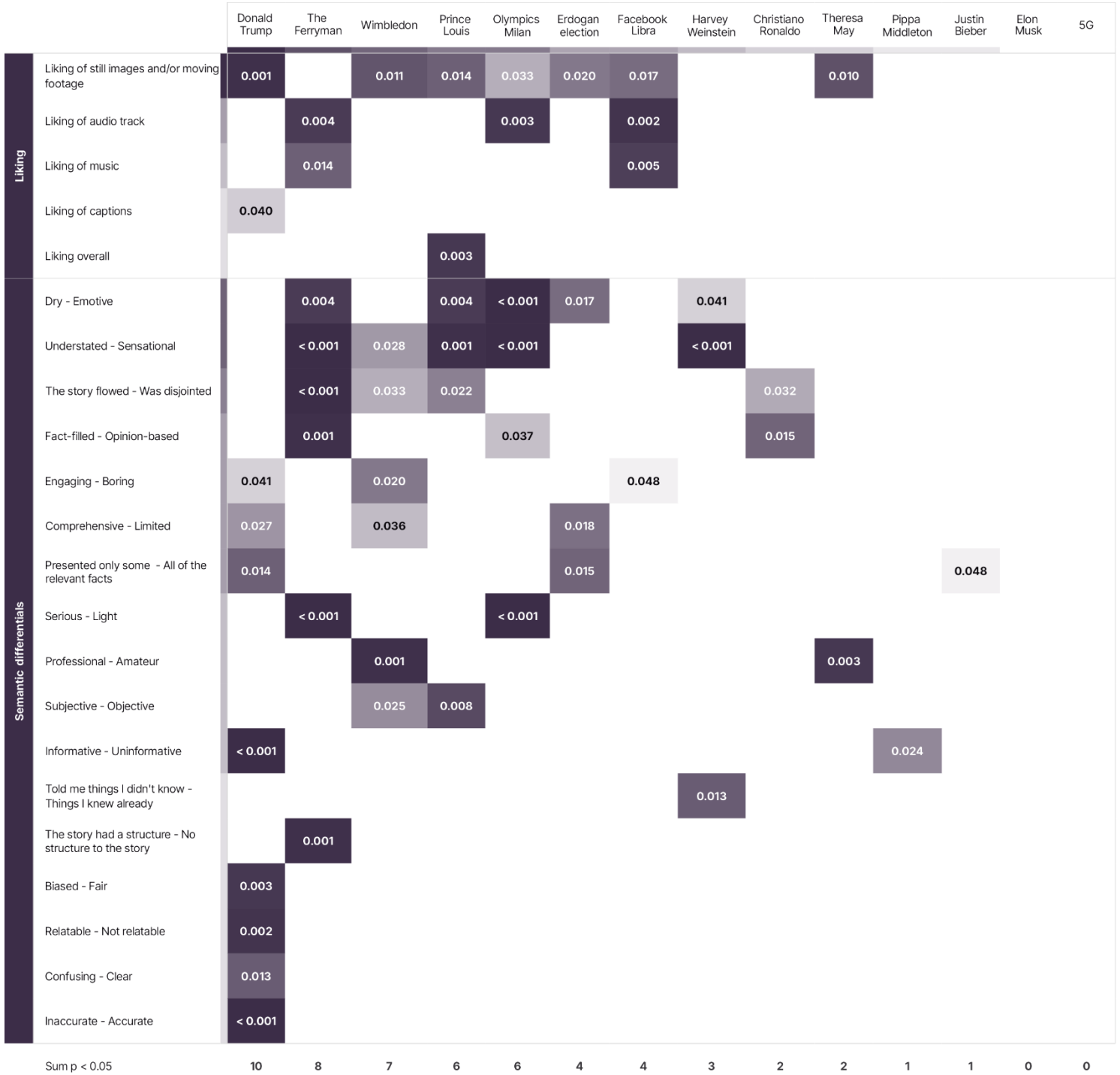

Figure 3: P-values for statistically significant differences <0.05 from one-way ANOVA tests conducted separately for each of 14 video topics. Dependent variables are continuous evaluation items; independent variable is video authorship (human-made, partly-automated, and highly-automated).

RQ2: How, if at all, do any differences found in RQ1 vary across the 14 story topics (e.g. Winter Olympics vs cryptocurrency vs Turkish election) included in our experiment?

Figure 3 and Table C in the supplemental material show the results of one-way ANOVAs within each of the 14 topic sets of videos, presenting the topic sets in order from those returning the largest number of contrasts—the set of Donald Trump videos, with statistically significant (at $p<0.05$) contrasts for ten evaluation items—to the smallest number (the Elon Musk and 5G sets with none). The patterns of where (i.e. for which questions/dependent variables) such differences are and aren't found broadly mirror the global tests, inevitably—but with at most half of the topic sets (seven) returning significant results for any one question/dependent variable, and sometimes (e.g. for "told me things I already knew") only one. The set of Trump videos is of particular interest in returning several significant differences between the differently authored videos (for presentation of relevant facts, bias/fairness, relatability, informativeness, and accuracy) that are diluted by the other topic sets of videos in a way that means there are no global significant differences on these criteria when topic sets are pooled. We note that increasing the sample sizes (in our study a reasonable but not terribly large n=300 per ANOVA) might identify more contrasts as "significant", but also that with multiple tests, we would, in the long run, expect two or three of our 54 statistically significant results (which constitute 18% of the total number of tests: 54 out of 22 variables x 14 video sets) to be false positives. Substantively, we conclude that differences found in RQ1 *do* vary by video topic. In terms of *how* they vary, we do not see any particular pattern by broad content area—it does not seem to be the case, for example, that videos on political subject matters systematically return more

significantly contrasting evaluations than those on celebrity news. There is a clear methodological take-home message, however: using a range of stories for a study of this type is advantageous. If, for example, we had used only the set of Trump videos, or only the set of 5G videos, for our experiment, we would have arrived at quite different conclusions.

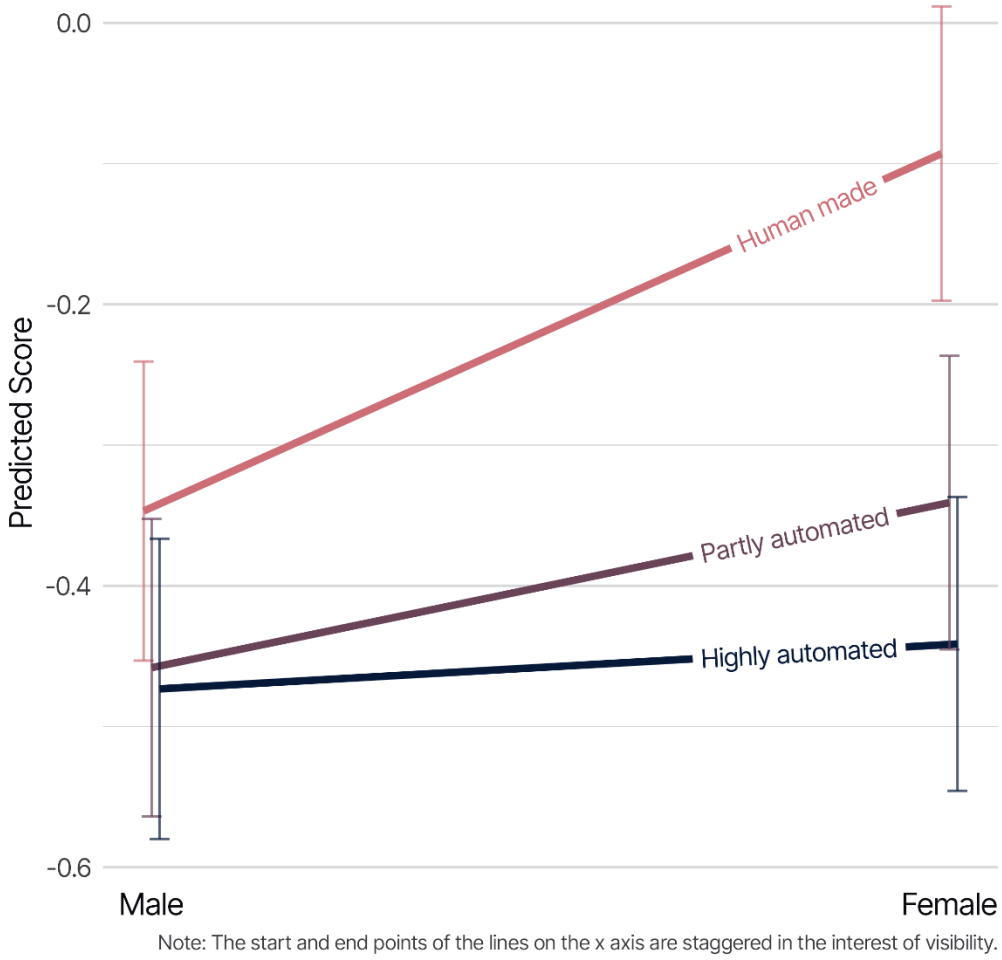

Figure 4. Illustration of interaction between video authors, gender of respondent, and evaluation of video as dry – emotive.

RQ3: How, if at all, do any differences found in RQ1 vary according to the gender, age, work status, region, and social grade of respondents?

To answer this question we ran multiple linear regressions with each of our evaluation variables as dependents, regressing on: the level of automation of the video; respondents' gender, age group, occupational status, region of residence, and social grade; and the video topic. We included interaction terms to test whether the way people evaluated the differently authored videos was moderated by any of these variables—in other words, whether any differences in perceptions of human-made versus partly or highly automated videos varied systematically between socio-demographic groups, and/or by video topic. Figure 4 illustrates such an association: male respondents tended to rate human-made and automated videos as similarly dry or emotive, but females tended to rate human-made videos as more emotive than the automated ones. Table 5 summarizes our results overall. The number of analyses makes it unfeasible to describe directions of association; instead, we simply highlight for which variables we found evidence of statistically significant associations between evaluation variable, video authorship, and socio-demographic variable. A letter "y" indicates that one or more individual coefficients was statistically significant (at $p<0.05$). Where these were found, the cell to the right gives the p-value for a joint Wald test for all coefficients for that interaction term. In several instances only one or two coefficients were significant, and the Wald test result was insignificant. In these cases, especially where the interaction involved a large number of categories, we would not want to read very much into just one or two parameter estimates being significant. Further analysis is needed to offer a detailed evaluation of these effects, noting also that with so many comparisons being tested here, some of the significant effects may be simply due to the play of chance. In that sense, it would be beneficial if future studies were able to maintain socio-

demographic variation with samples in order to add to the evidence base on this issue. Alongside these somewhat indeterminate results for socio-demographic variables, however, it is notable that interactions interactions involving the topic of video are far more often statistically significant. The results suggest that variation in video topic has a greater impact on variability of results than does socio-demographic variation, and that (all else being equal) maximizing the former rather than the latter would be more beneficial to future studies.

Table 5: Results of significance tests for whether socio-demographic and video topic variables moderate differential judgements of human-made and automated videos

	Gender	Age	Work status	Region	Social grade	Video topic
Liking (higher values = more liking)						
Liking of audio track						
Liking of music					y 0.120	y 0.243
Liking of still images and/or moving footage	y 0.078					
Liking of captions					y 0.089	
Liking overall	y 0.065			y 0.103	y 0.438	
Semantic differentials (lower values = first given word/phrase; higher values = second given word/phrase)						
told me things I didn't know - things I knew already	y 0.046	y 0.073	y 0.401			y 0.821
dry – emotive	y 0.108	y 0.257	y 0.265			y 0.000
serious – light				y 0.430		y 0.010
professional – amateur		y 0.306	y 0.340	y 0.495		y 0.240
fact-filled – opinion-based				y 0.172		y 0.016
engaging – boring	y 0.081					
comprehensive – limited			y 0.206		y 0.299	
story flowed – story was disjointed			y 0.694	y 0.256		y 0.001
understated – sensational				y 0.083	y 0.059	y 0.000
story had a structure – no structure to story	y 0.020			y 0.405		y 0.080
presented only some – all of the relevant facts			y 0.019	y 0.511		
biased – fair				y 0.367		y 0.813
subjective – objective						y 0.068
relatable – not relatable				y 0.736		y 0.210
informative – uninformative						y 0.053
confusing – clear						y 0.396
inaccurate – accurate			y 0.095	y 0.521		y 0.165

Legend: y indicates one or more significant ($p<0.05$) contrasts between categories; where y is given, the adjacent cell gives the p-value for a Wald test of all coefficients simultaneously (i.e. a test of the socio-demographic or video topic variable as a whole); p-values <0.05 are highlighted with grey fill.

Discussion and Conclusion

This study advances the scholarship on the perception of automated journalism in two main ways. Firstly, it explores an unresearched manifestation of the phenomenon, news video automation, a technology that is being used by national and international news publishers and that is attracting considerable investment, including from news agencies. Secondly, it employs a more rigorous methodology than has been seen in the literature to date. Consequently, our results have unprecedented validity and are also able to provide methodological guidance for future studies on the perception of automated journalism.

By using a 14 (video topic) x 3 (level of automation) between-subjects population-based online survey experiment, we were able to include an unusually large range of stimuli, enabling us to examine variations by story topic—potentially a confounding variable. Our large and diverse sample of respondents (n=4,200) also allowed us to explore whether differences in evaluations varied by five socio-demographic variables. The three levels of automation—high, partial, and none—that our independent variable contains is reflective of real-world practice where hybrid workflows are becoming more common (see, e.g., Renwick 2020), as reflected in our partially automated condition. In the acquisition and creation of our experimental stimuli we took great care to minimize or eliminate the influence of other potentially confounding variables, such as music, editing transitions, caption style and content, and video resolution. Furthermore, the major qualitative study (Koliska et al. 2021) and respondent-driven cognitive pretesting that we undertook to develop our dependent variables mean that the measures we use are highly appropriate to our object of study and respondents. We also followed survey methodological best practice in taking care to match the concept in the response options/scales to the concepts in the survey questions,

rather than using ‘agree-disagree’ responses. Finally, our survey scripting and data quality assurance procedures involved multiple steps to ensure that respondents had watched and listened to the video stimuli, that response order effects were minimized, and that suspicious or poor-quality responses were eliminated.

Our results show, for the first time, how news consumers evaluate highly and partially automated news videos against human-made equivalents. Although there was no significant difference in overall liking between the human-made and partly automated videos, there were statistically significant differences in 12 of the other 21 continuous evaluation items (see Figure 1 and Table C in the supplemental material). Those differences were not very large. It is a question for future consideration what would be considered theoretically to be a practically important difference. From our perspective, any signs that audiences evaluate automated and human-made videos differently at all is of note because they may alter the audience experience. The human-made videos tended to receive more positive ratings, for example on their professionalism; engaging nature; story flow; lack of sensationalism; and liking of audio, music, and images. However, there were no significant differences on other items, including accuracy, informativeness, objectivity, and bias.

These results share similarities with the results of previous studies on the perception of textual automated journalism that reported human-written news was evaluated as being of higher “quality” and more “readable” but no more “credible” than automated articles (Graefe and Bohlken 2020).

Some of the dimensions that made up the “quality” and “readability” criteria used in those previous studies—such as “well-written”, “clear”, “coherent”, “concise”, “professional”, “interesting”, “pleasing”, “enjoyable”—are the same as (“professional”) or somewhat equivalent to (“engaging”, “flowing”) the measures used in our study

under which the human-made videos were evaluated more highly. Although we have not, as previous studies have done, grouped our evaluative dimensions into broad composite criteria, our results overall suggest that automated news videos are somewhat less "watchable" than human-made equivalents, in a way that resembles how prior research found automated (textual) journalism to be less "readable".

Some of the dimensions that made up the "credibility" criterion used in those previous studies—specifically "accurate", "biased", and "objective"—were also reflected in items used in our study. And, like those studies, we found no significant differences between the evaluations of the human and automated videos on these dimensions. If we consider, for the purposes of comparison, these three dimensions to represent "credibility", then we too can say that we find no differences in how "credible" respondents thought human and automated videos to be. Our favoured explanation for why automated news videos were judged to be as credible as those made by humans is, as we have mentioned, because that criterion would naturally be judged on the basis of the content of the captions, which were human-written and did not vary by experimental condition. Therefore, in the context of our study we do not, as some other studies have suggested, see "accuracy, trustworthiness and objectiveness" as a "[strength] of … software-generated content" (Melin et al. 2018) but rather that the automated videos inherited their relative "credibility" from the human authors of their scripts.

We did not have the space in this article to fully explore the particular reasons why human-made videos scored more highly than automated videos on dimensions such as professionalism and engagement. However, our data shows that they were considered to have a greater variety of more relevant images/moving footage and to be more satisfactory in terms of the presence of a human voice-over. Exploring the role such

compositional/production characteristics play in variations in the evaluation of human and automated journalism is an area ripe for further study.

Finally, our results provide important methodological guidance for future studies on the perception of automated journalism. Firstly, by showing future research would benefit from maintaining socio-demographic variation within samples of respondents, and secondly, and more importantly, by showing that using a range of stories is important. A reduction in the sets of stories we used in our experiment would have led to quite different conclusions.

Notes

1. Defined as those who consume online news at least once a month.
2. As part of the survey panel provider's quality assurance, respondents are not told in advance the topic of the survey: this is to avoid the sample being skewed towards those with a particular interest in the subject matter.
3. For example, the items accurate-inaccurate, biased-fair, and subjective-objective consistently loaded together on a factor, but also with the item confusing-clear, which had a weaker cross-loading on a factor defined by the items relating to narrative structure (whether the story had a structure or not, and whether it flowed or was disjointed).

References

Carlson, Matt. 2015. "The Robotic Reporter: Automated Journalism and the Redefinition of Labor, Compositional Forms, and Journalistic Authority." *Digital Journalism* 3 (3): 416–431.

Charness, Gary, Gneezy, Uri, and Michael A. Kuhn. 2012. "Experimental Methods: Between-subject and Within-subject design." *Journal of Economic Behavior & Organization* (81)1: 1–8.

Clerwall, Christer. 2014. "Enter the Robot Journalist: Users' Perceptions of Automated Content." *Journalism Practice* 8 (5): 519–531.

Crunchbase. n.d.-1. https://www.crunchbase.com/organization/wibbitz/company_financials.

Crunchbase. n.d.-2. https://www.crunchbase.com/organization/whochit/company_financials.

Diakopoulos, Nicholas. 2019. *Automating the News: How Algorithms Are Rewriting the Media*. Harvard University Press.

Dörr, Konstantin Nicholas. 2016. "Mapping the Field of Algorithmic Journalism." *Digital Journalism* 4 (6): 700–722.

Fanta, Alexander. 2017. *Putting Europe's Robots on the Map: Automated Journalism in News Agencies*. Oxford: Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2017-09/Fanta%2C%20Putting%20Europe%E2%80%99s%20Robots%20on%20the%20Map.pdf.

Fletcher, Richard, Steve Schifferes, and Neil Thurman. 2020. "Building the 'Truthmeter': Training Algorithms to Help Journalists Assess the Credibility of Social Media Sources." *Convergence: The International Journal of Research into New Media Technologies* 26 (1): 19–34.

Graefe, Andreas, and 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. https://doi.org/10.17645/mac.v8i3.3019.

Graefe, Andreas, Mario Haim, and Nicholas Diakopoulos. 2017. "Should News Outlets Let their Readers Know that they Are Reading Automated Content? Effects of Algorithmic Transparency on Perceptions of Automated News." Poster presented at ICA 2017, San Diego, May.

Graefe, Andreas, Mario Haim, Bastian Haarmann, and Hans-Bernd Brosius. 2018. "Readers' Perception of Computer-generated News: Credibility, Expertise, and Readability." *Journalism* 19 (5): 595–610.

Jia, Chenyan. 2020. "Chinese Automated Journalism: A Comparison between Expectations and Perceived Quality." *International Journal of Communication* 14: 2611–2632.

Jung, Jaemin, Haeyeop Song, Youngju Kim, Hyunsuk Im, and 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.

Koliska, Michael, Thurman, Neil, Stares, Sally, and Jessica Kunert. 2021. "Exploring audience perceptions of, and preferences for, online news videos." *Journalism Studies*, 22(9): 1161–1180.

Krosnick, Jon, and Stanley Presser. 2010. "Question and Questionnaire Design." In *Handbook of Survey Research*, edited by Peter V. Marsden and James D. Wright, 2nd ed., 263–315. Emerald.

Kunert, Jessica, and Neil Thurman. 2019. "The Form of Content Personalisation at Mainstream, Transatlantic News Outlets: 2010–2016." *Journalism Practice* 13 (7): 759–780.

Li, Yifei, Mingyang Yu, and Sichen Li. 2022. "Technology or Content: Which Factor is More Important in People's Evaluation of Artificial Intelligence News?" *Telematics and Informatics Reports*, 9: 1–7.

Liu, Bingjie, and Lewen Wei. 2019. "Machine Authorship *In Situ*: Effect of News Organization and News Genre on News Credibility." *Digital Journalism* 7 (5): 635–657.

Logg, Jennifer M. 2017. "Theory of Machine: When Do People Rely on Algorithms?" Harvard Business School Working Paper No. 17-086. https://dash.harvard.edu/handle/1/31677474.

Logg, Jennifer M., Julia A. Minson, and Don A. Moore. 2019. "Algorithm Appreciation: People Prefer Algorithmic to Human Judgment." *Organizational Behavior and Human Decision Processes* 151: 90–103.

Maes, A., Nicole Ummelen, and Hans Hoeken. 1996. *Instructieve teksten. Analyse, ontwerp en evaluatie*. Bussum: Coutinho.

Melin, Magnus, Asta Bäck, Caj Södergård, Myriam Munezero, Leo Leppänen, and Hannu Toivonen. 2018. "No Landslide for the Human Journalist: An Empirical Study of Computer-generated Election News in Finland." *IEEE Access* 6: 43356–43367.

Mutz, Diana C. 2011. "Population-based Survey Experiments: A Hybrid Methodology for the Social Sciences." In *Population-Based Survey Experiments*. Princeton: Princeton University Press.

Newman, Nic, Richard Fletcher, David A. L. Levy, and Rasmus Kleis Nielsen, eds. 2016. *Reuters Institute Digital News Report 2016*. Oxford.

Newman, Nic, Richard Fletcher, Anne Schulz, Simge Andı, and Rasmus Kleis Nielsen, eds. 2020. *Reuters Institute Digital News Report 2020*. Oxford.

Presser, Stanley, and Johnny Blair. 1994. "Survey Pretesting: Do Different Methods Produce Different Results?" *Sociological Methodology* 24: 73–104.

Renwick, Alan. 2020. "Rethinking the Economics of Local Public Interest News – Combining Journalists, Data and Automation to Make Quality Content Affordable at a Granular Level." Paper presented at Challenges of Journalism in the 21st Century – Automated and AI Journalism conference, Charles University, Prague, 24 September. https://www.youtube.com/watch?v=0X5804a7c7g&t=1s.

Sundar, S. Shyam. 1999. "Exploring Receivers' Criteria for Perception of Print and Online News." *Journalism & Mass Communication Quarterly* 76 (2): 373–386.

Tandoc, Edson C. Jr, Lim Jia Yao, and Shangyuan Wu. 2020. "Man vs. Machine? The Impact of Algorithm Authorship on News Credibility." *Digital Journalism* 8 (4): 548–562.

Thäsler-Kordonouri Sina, and Kurt Barling. 2023. "Automated Journalism in UK Local Newsrooms: Attitudes, Integration, Impact." *Journalism Practice*, DOI: 10.1080/17512786.2023.2184413.

Thurman, Neil. 2018. "Social Media, Surveillance, and News Work: On the Apps Promising Journalists a 'Crystal Ball'." *Digital Journalism* 6 (1): 76–97.

Thurman, Neil. 2019. "Computational Journalism." In *The Handbook of Journalism Studies*, edited by Karin Wahl-Jorgensen and Thomas Hanitzsch. 2nd ed. New York: Routledge.

Thurman, Neil, Konstantin Dörr, and Jessica Kunert. 2017. "When Reporters Get Hands-on with Robo-writing: Professionals Consider Automated Journalism's Capabilities and Consequences." *Digital Journalism* 5 (10): 1240–1259.

Thurman, Neil, Stares, Sally, and Michael Koliska. 2023. "Replication data for: Audience evaluations of online news videos made with various levels of automation, and none: A population-based survey experiment." https://figshare.com/s/c33a160473152e3e18b4.

Van der Kaa, Hille, and Emiel Krahmer. 2014. “Journalist versus News Consumer: The Perceived Credibility of Machine Written News.” In Proceedings of the Computation+Journalism conference New York. https://pure.uvt.nl/portal/files/4314960/c.

Van der Lee, Chris, Bart Verduijn, Emiel Krahmer, and 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, 962–972.

Waddell, T. Franklin. 2018. “A Robot Wrote This? How Perceived Machine Authorship Affects News Credibility.” *Digital Journalism* 6 (2): 236–255.

Wibbitz n.d. https://www.wibbitz.com/solution-wavi/.

Wimbledon n.d. “Wimbledon's Cognitive Highlights.” https://www.wimbledon.com/en_GB/aboutwimbledon/wimbledons_cognitive_highlights.html.

Wochit n.d. https://www.wochit.com/products/wave/.

Wölker, Anja, and Thomas E. Powell. 2021. “Algorithms in the Newsroom? News Readers’ Perceived Credibility and Selection of Automated Journalism.” *Journalism* 22 (1): 86–103.

Zheng, Yue, Bu Zhong, and Fan Yang. 2018. “When Algorithms Meet Journalism: The User Perception to Automated News in a Cross-cultural Context.” *Computers in Human Behavior* 86: 266–275

Supplemental Material

Table A. News videos used as stimuli in the survey experiment.

	Story		Videos			Editing undertaken on "highly automated" video to create "partly automated" video	Category
			Human-made	Partly automated	Highly automated		
1	Pippa Middleton, sister of the Duchess of Cambridge, announces she is expecting a baby.	Has music?	Yes	Yes	Yes	14/22 scenes edited = 64%	Royalty
		Has background sound?	No	No	No		
		Uses mix of still and moving images?	Yes	Yes	Yes		
		Length (secs)	59	70	70		
2	A review of Cristiano Ronaldo's career in 60 seconds.	Has music?	Yes	Yes	Yes	15/25 scenes edited = 60%	Sport
		Has background sound?	No	No	No		
		Uses mix of still and moving images?	Yes	Only Stills	Only Stills		
		Length (secs)	62	64	67		
3	Elon Musk apologizes for making accusations against British cave diver.	Has music?	Yes	Yes	Yes	9/17 scenes edited = 53%	Business
		Has background sound?	Yes	No	No		
		Uses mix of still and moving images?	Yes	Only Stills	Only Stills		
		Length (secs)	54	87	87		
4	Recep Tayyip Erdogan wins Turkey's presidential election.	Has music?	Yes	Yes	Yes	9/14 scenes edited = 64%	Politics
		Has background sound?	No	No	No		
		Uses mix of still and moving images?	All Moving	Only Stills	Only Stills		
		Length (secs)	41	52	53		

5	Harvey Weinstein pleads not guilty to rape in New York court hearing.	Has music?	No	Yes	Yes	4/19 scenes edited = 21%	Crime / celebrity
		Has background sound?	Yes	No	No		
		Uses mix of still and moving images?	All Moving	Only Stills	Only Stills		
		Length (secs)	60	64	65		
6	Canadian singer Justin Bieber announces his engagement to Hailey Baldwin.	Has music?	No	Yes	Yes	7/13 scenes edited = 54%	Celebrity
		Has background sound?	Yes	No	No		
		Uses mix of still and moving images?	Yes	Yes	Yes		
		Length (secs)	37	47	47		
7	Prince Louis of Cambridge is christened.	Has music?	No	Yes	Yes	8/14 scenes edited = 57%	Royalty
		Has background sound?	Yes	No	No		
		Uses mix of still and moving images?	All Moving	Yes	Only Stills		
		Length (secs)	39	49	49		
8	Donald Trump's commemorative star on the Hollywood Walk of Fame is destroyed.	Has music?	No	Yes	Yes	11/17 scenes edited = 65%	Politics / crime / celebrity
		Has background sound?	Yes	No	No		
		Uses mix of still and moving images?	All Moving	Only Stills	Yes		
		Length (secs)	55	53	53		
9	Milan–Cortina is chosen to host the 2026 Winter Olympic and Paralympic Games.	Has music?	No	Yes	Yes	6/11 scenes edited = 55%	Sport
		Has background sound?	Yes	No	No		
		Uses mix of still and moving images?	All Moving	Only Stills	Only Stills		
		Length (secs)	45	38	38		

10	An explainer video on 5G wireless mobile technology.	Has music?	Yes	Yes	Yes	22/41 scenes edited = 54%	Technology
		Has background sound?	No	No	No		
		Uses mix of still and moving images?	All Moving	Yes	Yes		
		Length (secs)	90	126	126		
11	An explainer video on Facebook's cryptocurrency, Libra.	Has music?	Yes	Yes	Yes	22/32 scenes edited = 69%	Technology
		Has background sound?	No	No	No		
		Uses mix of still and moving images?	All Moving	Yes	Yes		
		Length (secs)	87	99	99		
12	Northern Irish play The Ferryman wins best play at the Tony Awards.	Has music?	No	Yes	Yes	7/16 scenes edited = 44%	Culture
		Has background sound?	Yes	No	No		
		Uses mix of still and moving images?	All Moving	Yes	Yes		
		Length (secs)	48	50	50		
13	Theresa May announces her decision to step down as Conservative Party leader and UK Prime Minister.	Has music?	Yes	Yes	Yes	14/23 scenes edited = 61%	Politics
		Has background sound?	Yes	No	No		
		Uses mix of still and moving images?	Yes	Only Stills	Only Stills		
		Length (secs)	59	76	75		
14	A video highlighting numerical facts about the Wimbledon tennis championships.	Has music?	Yes	Yes	Yes	9/17 scenes edited = 53%	Sport
		Has background sound?	Yes	No	No		
		Uses mix of still and moving images?	Yes	Only Stills	Only Stills		
		Length (secs)	43	50	50		

Survey questions

Question stem in regular bold font.
Response options in regular font.
[Notes (not seen by respondent) in regular italicised font within square brackets.]

What is your date of birth? *[Year and month selected from drop-down menu]*

Which of the following describes how you think of yourself?
1 Male
2 Female
3 In another way
4 Prefer not to say

Where do you live?
Please note: This question may be considered personal. We would like to remind you that your participation is strictly voluntary and that your responses are used for research purposes only. The answers that you provide will be presented in aggregate form and none of them will be linked back to you in any way. All data will be collected and processed in adherence to the Market Research Society's Code of Conduct and the General Data Protection Regulation (GDPR).
What is your Postcode? *[Postcode, Postal Town, or Prefer Not to Answer given]*

Which of the following sources, if any, do you regularly use for news? By regularly we mean at least once a month. Please think about all the ways and times you use news. This could be at home or out and about, and could be on any device such as TV, any radio, newspapers, computer, mobile phone, tablet. Select all that apply.
1. TV
2. Radio
3. Newspapers
4. Newspaper websites/apps
5. Other news websites/apps
6. News on social media sites/apps
7. None of these [single code]
8. Don't know [single code]

[Screening question: survey closes if 4, 5 or 6 is not selected]

Thank you for taking part in this study, which is being conducted by Ipsos MORI on behalf of researchers at [REDACTED]. We are researching people's reactions to online news videos.
In this survey we will ask you to watch one short news video and then answer questions about it. At the end there are a few questions about you.
All of your answers will be treated in the strictest confidence. They will be reported as aggregated statistics, you will not be linked to your responses and you will not be identified.
Please indicate whether you are happy to proceed to the survey. If you start the survey but do not wish to continue, you can stop answering questions at any time by closing the browser window.
1 I am happy to participate in the survey
2 I do not wish to participate

[Survey closes if 2 is selected]

Please click the button below to hear a sound. Please make sure the volume is turned on so you can hear the sound. [Sound button placed below: sound is of a cow mooing.] **What sound did you hear? Please select one answer only.**
1 A cow mooing
2 A bird tweeting
3 A dog barking
4 A lion roaring

[Survey closes if 2, 3 or 4 is selected]

On the next screen you will watch a video. When you are ready, click the button to advance to the next screen. Please click the play button to watch the video. You can watch it more than once if you like. The video will be included at the bottom of some of the other question pages so that you can watch it again when answering the survey questions, if you wish.

[Respondents unable to advance to the next screen for the period of time corresponding to the length of the video. This is to ensure that they watched the video in its entirety and did not fast-forward the video or skip it altogether.]

Did you hear the video's sound when you watched it? Please select one answer only.

1 Yes, clearly
2 Yes, partially
3 No, I couldn't hear it

[Survey closes if 2 or 3 is selected]

Did the video play smoothly? Please select one answer only.

1 Yes, it played smoothly
2 There were a few pauses or glitches
3 There were many pauses or glitches
4 The video didn't play

[Survey closes if 2, 3 or 4 is selected]

First, we'd like to ask you for some of your general reactions to the video that you just watched.
How much did you like or dislike the video? Please position the slider on or between the numbers to show how much you liked or disliked it. The higher the number, the more you liked it, the lower the number, the more you disliked it.

[For all 'liking' items, such as this, scale numbers not displayed to respondents. Slider can be positioned at any point at 0.1 unit intervals]

I disliked it very much — I neither liked nor disliked it — I liked it very much

You will now be shown several pairs of statements. For each pair of statements, please move the slider towards the statement that best describes your personal opinion of the video. The closer you put the slider to one end or the other the more you should think the video fits the statement at that end. If you feel neutral, or agree with each statement equally, leave the slider in the middle.

[Scales coded -3 to +3; numbers not displayed to respondents. Slider snapped to integer values.]

inaccurate	[Slider]	accurate
biased	[Slider]	fair
fact-filled	[Slider]	opinion-based
subjective	[Slider]	objective
told me things I didn't know	[Slider]	told me things I knew already
presented only some of the relevant facts	[Slider]	presented all of the relevant facts
the story had a structure	[Slider]	there was no structure to the story
the story flowed	[Slider]	the story was disjointed

Here are some more scales. Again, for each pair of statements, please move the slider towards the statement that best describes your personal opinion of the video. The closer you put the slider to one end or the other the more you should think the video fits the statement at that end. If you feel neutral, or agree with each statement equally, leave the slider in the middle.

relatable	[Slider]	not relatable
informative	[Slider]	uninformative
professional	[Slider]	amateur
dry	[Slider]	emotive
serious	[Slider]	light
engaging	[Slider]	boring
understated	[Slider]	sensational
confusing	[Slider]	clear
comprehensive	[Slider]	limited

Now thinking about the overall length of the video, would you have liked it to be shorter, longer or did you think it was about right for the story? Please select one answer only.

1 I would have liked it to be shorter
2 Its length was about right for the story
3 I would have liked it to be longer
98 I don't have an opinion

[1–3 scale reversed for alternate respondents to mitigate primacy effects; code 98 anchored to the end of the scale]

Now for some questions about the use of captions (the words on the screen).
Firstly, how much did you like or dislike the use of captions in general? Please position the slider on or between the numbers to indicate your view.

Disliked very much — Neither liked nor disliked — Liked very much

And now some more specific questions about the captions.
For the characteristic stated below, please select one response that best represents your personal opinion on the video overall. Please select one answer only.

Readability of the captions	Difficult to follow overall	In some places difficult, in other places easy to follow	Easy to follow overall	Don't have an opinion
Size of the text	Too small	About the right size	Too big	Don't have an opinion
Amount of text	Not enough	About the right amount	Too much	Don't have an opinion
Speed of movement/transition of captions	Too slow	About the right speed	Too fast	Don't have an opinion

And now for some questions about the other visual elements of the video.
On balance, how much did you like or dislike the still images and/or moving footage that was shown within the video? Please position the slider on or between the numbers to indicate your opinion.

Disliked very much — Neither liked nor disliked — Liked very much

How relevant did you think the still images and/or clips of moving footage were to the story? Please select one answer only.

5 They all related well to the story
4 Most related well to the story
3 There was a roughly even mixture of related and unrelated ones
2 Only some related well to the story
1 None related well to the story
98 Don't know

[5–1 scale reversed for alternate respondents; code 98 anchored to the end of the scale]

How would you rate the quality of the still images and/or moving footage? By quality we mean things like how they were framed, their lighting and so on. Please select one answer only.

5 High quality
4 Good quality
3 Of mixed quality
2 Poor quality
1 Very poor quality

[5–1 scale reversed for alternate respondents]

How well or badly did the still images and/or moving footage match the words in the captions? Please select one answer only.
5 Very well
4 Well
3 Satisfactorily
2 Not well
1 Not at all well
98 Don't know
[5–1 scale reversed for alternate respondents; code 98 anchored to the end of the scale]

And what about the variety of still images and/or clips of moving footage within the video? Please select one answer only.
1 They were too similar to each other
2 There was a good balance between similar and varied ones
3 They were too varied
[1–3 scale reversed for alternate respondents]

How much did you like the video's audio track? Please position the slider on or between the numbers to indicate your opinion.

Disliked very much Neither liked nor disliked Liked very much

Online news videos sometimes contain music, sounds recorded on location (e.g. voices, bustling crowds etc.), a voiceover from a journalist or a mixture of these elements. Would you have liked to have heard more of any of these elements in the video you watched? *Please select all that apply*.
1 Location sound
2 Music
3 A voiceover from a journalist
4 It was fine as it was

[The following question asked only to those who have seen a video with location sound]
Which of the following statements best matches your opinion on the use of location sound (e.g. voices, bustling crowds etc.)? Please select one answer only.
The use of location sound…
3 Enhanced my experience of the news story
2 Had no impact on my experience of the news story
1 Worsened my experience of the news story
98 I did not hear any location sound
[3–1 scale reversed for alternate respondents; code 98 anchored to the end of the scale]

[The following three questions asked only to those who have seen a video with music]
Now we'd like to ask for your views on the music in the video.

How much did you like or dislike the music in the video? Please position the slider on or between the numbers to indicate your opinion.

Disliked very much Neither liked nor disliked Liked very much

Which of the following statements best matches your opinion on the use of music? Please select one answer only.
The music...
1 Was suited to the news story
2 Was not suited to the news story
98 Don't know

And which of the following statements best matches your opinion on the use of music? Please select one answer only.
The use of music...
3 Enhanced my experience of the news story
2 Had no impact on my experience of the news story
1 Worsened my experience of the news story

These days there are computer programmes that can automatically generate short news videos based on a text story provided by a journalist. Sometimes these videos are then further edited by a journalist before being published. How do you think the video you watched was made? Please select one answer only.
1 Completely made by a journalist (human made)
2 Mostly made automatically by computer software and then further edited by a human journalist (partly automated)
3 Mostly made automatically by computer software (highly automated)
[1–3 scale reversed for alternate respondents]

What was the video you just watched about? Please select one answer only.
1 Pippa Middleton expecting a baby
2 Cristiano Ronaldo
3 Elon Musk's apology to a British cave diver
4 Turkey's presidential election
5 Harvey Weinstein's New York court hearing
6 Justin Bieber's engagement to Hailey Baldwin
7 Christening of Prince Louis of Cambridge
8 Donald Trump's star on the Hollywood Walk of Fame
9 2026 Winter Olympic Games
10 5G wireless mobile technology
11 Facebook's cryptocurrency, Libra
12 The Ferryman winning best play at the Tony Awards
13 Theresa May
14 The Wimbledon tennis championships
98 Don't know
99 None of these
[1–14 order randomised; codes 98 and 99 anchored to the end of the scale. This was an attention check question, and if respondents answered incorrectly the survey would close for them, and their answers to other questions would not be included in the final sample. They were not able to go back to watch their designated video again before answering this question.]

Which of the following devices did you watch this video on? Please select one answer only.
1 Laptop or desktop computer
2 Smartphone
3 Tablet computer (e.g. iPad)
4 Other: please specify

And finally, we'd like to ask you a few questions about your news consumption.
Firstly, how interested, if at all, would you say you are in news in general? Please select one answer only.
1 Very interested
2 Fairly interested
3 Not very interested
4 Not at all interested
98 Don't know
[1–4 scale reversed for alternate respondents; code 98 anchored to the end of the scale.]

How interested are you in the following types of news? Please select one answer per option.
1 International news
2 Political news
3 News about my region, city or town
4 Business and economic news
5 Entertainment and culture
6 News about celebrities including the Royal Family
7 Lifestyle news (e.g. food, fashion, travel, cooking, wellness)

8 Health or education news
9 Sports news
10 Science and technology news
11 News about crime, justice and security
12 News about the environment
13 Weird news (e.g. funny, bizarre, quirky)
14 Other

[Order of items randomised but with 14 anchored at the end of the list. Response options given below. 1–4 scale reversed for alternate respondents; code 98 anchored to the end of the scale.]

1 Very interested
2 Fairly interested
3 Not very interested
4 Not at all interested
98 Don't know

Which of the following brands, if any, have you used to access news <u>online</u> in the last week (via websites, apps, social media, and other forms of Internet access)? Please select all that apply.

1 BBC News online
2 CNN.com
3 ITV News online
4 Sky News online
5 C4 News online
6 C5 News online

7 BuzzFeed News
8 Vice News
9 HuffPost (Huffington Post)
10 MSN News
11 Yahoo! News
12 The Lad Bible news

13 Guardian online
14 Times online
15 Telegraph online
16 Independent/ i100 online
17 inews.co.uk
18 Financial Times online
19 Economist online
20 New York Times online

21 Mail online
22 The Sun online
23 Mirror online
24 Express online

25 Metro online
26 Scotsman online
27 The Herald online
28 Daily Record online
29 Wales online
30 The Irish News online
31 Belfast Newsletter online
32 Belfast Telegraph online
33 London Evening Standard online

34 Other regional or local newspaper website
35 Other online sites from outside the UK

98 Don't know
99 None of these

[Order of blocks, and items within blocks, randomised, but with 'other' block (34 and 35) and codes 98 and 99 anchored at the end of the list.]

Which of the following brands, if any, have you used to access news <u>offline</u> in the last week (via TV, radio, print, and other traditional media)? Please select all that apply.

1 BBC TV News
2 ITV News
3 Sky News
4 C4 News
5 C5 News
6 CNN
7 Russia Today

8 BBC Radio News
9 Commercial radio news

10 Daily Mail / Mail on Sunday
11 The Sun / Sun on Sunday
12 Daily Mirror / Sunday Mirror / Sunday People
13 The Express / Sunday Express
14 Daily Star / Star Sunday

15 The Times / Sunday Times
16 The Daily Telegraph / Sunday Telegraph
17 The Guardian / Observer
18 The 'i'
19 The Financial Times
20 Metro
21 Economist

22 The Scotsman
23 The Herald
24 Daily Record
25 Western Mail / Wales On Sunday
26 The Irish News
27 Belfast Newsletter
28 Belfast Telegraph
29 London Evening Standard

30 A regional or local newspaper
31 Other newspapers or broadcast news channels from outside the UK

98 Don't know
99 None of these

[Order of blocks, and items within blocks, randomised, but with 'other' block (30 and 31) and codes 98 and 99 anchored at the end of the list.]

Table B. Composition of survey respondents (n=4,200).

		%
Gender	Male	49
	Female	51
Age	18–34	32
	35–54	39
	55–75	30
Social grade	Upper middle class	4
	Middle class	41
	Lower middle class	34
	Skilled working class	8
	Working class	10
	Lower level of subsistence	9

Note: Percentages may not sum to 100 due to rounding.