Auditing the Information Quality of News-Related Queries on the Alexa Voice Assistant

HENRY KUDZANAI DAMBANEMUYA, Northwestern University, USA
NICHOLAS DIAKOPOULOS, Northwestern University, USA

Smart speakers are becoming increasingly ubiquitous in society and are now used for satisfying a variety of information needs, from asking about the weather or traffic to accessing the latest breaking news information. Their growing use for news and information consumption presents new questions related to the quality, source diversity, and comprehensiveness of the news-related information they convey. These questions have significant implications for voice assistant technologies acting as algorithmic information intermediaries, but systematic information quality audits have not yet been undertaken. To address this gap, we develop a methodological approach for evaluating information quality in voice assistants for news-related queries. We demonstrate the approach on the Amazon Alexa voice assistant, first characterising Alexa's performance in terms of response relevance, accuracy, and timeliness, and then further elaborating analyses of information quality based on query phrasing, news category, and information provenance. We discuss the implications of our findings for future audits of information quality on voice assistants and for the consumption of news information via such algorithmic intermediaries more broadly.

CCS Concepts: • **General and reference** → **Evaluation**; • **Computing methodologies** → *Model development and analysis.*

Additional Key Words and Phrases: algorithmic accountability, voice assistants, information quality, audit framework

ACM Reference Format:
Henry Kudzanai Dambanemuya and Nicholas Diakopoulos. 2021. Auditing the Information Quality of News-Related Queries on the Alexa Voice Assistant. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 83 (April 2021), 21 pages. https://doi.org/10.1145/3449157

1 INTRODUCTION

As of April 2020 approximately 1-in-4 American adults (~60 million) owned smart speaker devices (e.g. Amazon Alexa, Google Home, etc.) according to the latest NPR Smart Audio Report [57]. In NPR's earlier survey of more than 800 smart speaker owners in mid-2019, 74% reported that they relied on smart speakers and other voice assistants to answer general questions and 42% said they often sought news information [56]. While search engines and social media platforms are still the dominant algorithmic intermediaries for news information [15, 47], the growing use of smart speakers for news and information presents new questions related to the quality, source diversity, and comprehensiveness of information conveyed by these devices. How do smart speakers respond when asked specific questions about the news? And what information sources do these devices use to

Authors' addresses: Henry Kudzanai Dambanemuya, Northwestern University, USA, 2240 Campus Drive, 1-147, Evanston, Illinois, 60208, hdambane@u.northwestern.edu; Nicholas Diakopoulos, Northwestern University, USA, 2240 Campus Drive, 2-254, Evanston, Illinois, 60208, nad@northwestern.edu.

2573-0142/2021/4-ART83 $15.00
https://doi.org/10.1145/3449157

respond to news-related queries? These are particularly salient questions in light of the challenges media platforms have had in curating quality information and combating disinformation [20]. While other algorithmic intermediaries for news information, such as Google Search [43, 53, 58, 63], Google News [22], and Apple News [4] have recently been audited, similar audits examining the sources and quality of information curated by smart speaker devices have not yet been undertaken.

In order to address this gap, *we present a data-driven methodological approach for evaluating the information quality of voice assistants which is tailored to both users' information seeking behaviours and the specific capabilities and input modalities of these systems.* We focus on information quality because it is an issue that information intermediaries continue to struggle with [9, 42] and is particularly important for voice assistants because typically only one voice query result is provided instead of a set of results to choose from, thus raising the stakes for the quality of that one response. We tailor our approach to the input modalities of voice assistants by relying on crowd sourced question phrasings that people commonly use to elicit answers to general questions related to news information from voice assistants. Through a combination of automated and crowd-based methods we address the complexities of auditing voice assistants due to errors of speech synthesis and transcription algorithms. These systems depend on such capabilities to accurately listen, process, and respond to users' voice queries, which may exhibit a challenging array of phonological and prosodic variations.

We then demonstrate the application of our approach to the Amazon Alexa voice assistant on the Echo Plus smart speaker device and report the findings of our audit of information quality. We focus on Alexa here as it largely dominates the market for smart speakers (74% net market share in the UK and 63% in the US [48]). In undertaking this work we identify and discuss key information quality issues of Amazon's Alexa, elaborating analyses based on query phrasing, news category, and information provenance. This helps to demonstrate a technological gap in the ability of the Alexa voice assistant to provide timely and reliable news information about important societal and political issues, and suggests areas in which future generations of smart speakers and voice assistants might be improved. Our aim in this work is not to characterise the overall state of smart speaker information quality per se, since it is constantly evolving, but to highlight important challenges to voice-based algorithmic audits and elaborate a methodological approach for auditing information quality on voice assistants that can be used over time to characterise, compare, and track these systems.

2 RELATED WORK

In the following subsections we elaborate existing literature in two broad areas: (1) studies of smart speakers focusing on patterns of use and evaluations of user experience and (2) broader studies of algorithmic intermediaries for news media. Within these two areas of study, we articulate a research gap related to evaluating the quality of algorithmically-curated news-related information in smart speakers, which the current study aims to fill.

2.1 Studies of Smart Speakers

Recent studies of smart speakers have focused on topics such as the potential security risks and privacy concerns of the devices [11, 19, 34], the effects of uncertainty and capability expectations on users' intrinsic motivations for virtual assistant interaction [36], biases related to performance for specific languages or accents of different groups of people [38], the general use of conversational user interface techniques [39, 41, 60], and the role of conversational grounding in supporting user interactions with voice assistants [10].

Studies examining smart speaker patterns of use indicate a variety of tasks afforded by voice assistants. In one study focusing on children's use of conversational agents, a two-week in-home

deployment of Google's voice controlled digital assistant examined how young children use voice assistants to ask questions and find information over an extended period of time [41]. The study shows that children think about these devices as information sources for topics that they are curious about such as science and technology (24% of all questions) and culture (21%). These two categories account for a large proportion of questions asked and include questions about animals, plants, and nature as well as questions about celebrities and popular culture. Another study that investigates the experiences of households living with the Amazon Echo smart speaker was based on the history logs of Alexa and in-home contextual interviews with Alexa owners. The study showed differences in daily usage patterns across households ranging from listening to music and setting timers to asking for weather information [60].

A usability test of multiple speech-based natural user interfaces including Alexa, Siri, Cortana, and Google revealed several opportunities for improvement as the voice assistants performed differently across different task categories such as shopping, travel, entertainment, and administrative tasks [40]. Given the large variability of tasks afforded by voice assistants, Jiang et al developed a model for evaluating user experiences across different tasks in voice assistants on mobile devices [25]. Their model uses implicit user feedback in the form of user interaction sequences and generic features to predict users' satisfaction with intelligent voice assistants.

In contrast to these previous studies which look across various tasks and patterns of use, our study focuses on the specific task of seeking news-related information from smart speakers. In particular, we are interested in the quality of the information provided for news-related queries, an area of inquiry which has not yet received sufficient scholarly attention (except for the preliminary analysis reported in [14]), but which is important given the role that smart speakers are increasingly playing as algorithmic intermediaries for news-related information [48].

2.2 Audit Studies of Algorithmic Information Intermediaries

Algorithmic information intermediaries provide the means for both information *filtering* (i.e. inclusion and exclusion of information) and information *prioritisation* (i.e. drawing more attention to some subset of information) [7]. A growing body of research considers the role of such algorithmic intermediaries in curating news information and directing human attention in search engines, news apps, and social media feeds. This includes the auditing of informational biases that may result from the data or algorithms driving these systems [32], which may in turn result in increased or decreased exposure to different (1) sources of information, and (2) types of content.

Studies of Google search have shown that there is often a high concentration of attention given to a relatively narrow range of information sources, which may impact the diversity of information that users are ultimately exposed to [16, 63], and lead to reliance on particular sources [18, 65]. Further analysis of traffic to news sites from Google shows high levels of popularity bias, the tendency of an application to generate traffic to already popular websites and reinforce exposure to those popular news sites [50]. Other research demonstrates the importance of the choice of topic and query phrasing when auditing search information intermediaries [33, 63], which informs the audit method we develop in this work.

Algorithmic curation can also impact the nature of content that gains exposure. For instance, an audit of the algorithmically curated trending section on Apple News showed that it tended to surface more soft news in comparison to the manually curated section of the app [4]. Of particular interest in this work is how the inclusion, exclusion, and prioritisation of news information raises issues of information reliability and credibility [44, 59]. To measure the impact of algorithmic intermediaries on the attention given to content, Jurgens and Stark (2017) developed a model that captures and differentiates independent effects of filtering and sorting [26]. Their model, evaluated on Reddit, shows that information intermediaries shape users' information environments

as changes in various sections of the website included on the home page by default led to large changes in attention to content.

To a large extent, prior studies of algorithmic information intermediaries tend to focus on the source and diversity of results without evaluating the quality of the informational content itself. We believe that evaluating the quality of algorithmically-curated content is important, now more that ever, given the growing societal challenges of mis- and disinformation and the role algorithmic intermediaries play in access to and amplification of information. We thus contribute to the growing body of literature on audits of information intermediaries, but explicitly focus on information quality, providing a framework to audit algorithmic intermediaries for information quality that is tailored to both users' information seeking behaviours and the specific capabilities and input modalities of smart speaker systems.

3 HOW DOES ALEXA WORK?

The voice assistant, Alexa, can perform several tasks ranging from information retrieval, providing information, delivering news, telling the weather, controlling other smart devices, and allowing users to order products from its parent company, Amazon. Of interest to this study is how Alexa provides information and delivers news to its users. Users begin by saying a wake word such as "Echo", "Alexa", or "Amazon" to activate the voice assistant, followed by a voice query – in our case, a request for specific information or news, such as "Alexa, what is happening with Ilhan Omar today?" The voice query is then transcribed using speech recognition, machine learning, and natural language understanding algorithms. The transcription is sent as a request to the Alexa service which provides a text response that is converted to an audio response through a text-to-speech algorithm. In contrast to text-based search in most web-based algorithmic intermediaries, these key processes shown in Figure 1 are unique to voice interfaces and necessitate an evaluation of the voice assistant's speech synthesis and transcription capabilities (see Section 4.3) in order to qualify the impact of those system elements on the audit of information quality.

Fig. 1. Information architecture of Alexa. Source: [1]

The Alexa voice assistant, like web search engines, selects some information while filtering out a lot of other information in its response. However, unlike search engine results pages (SERPs) where a ranked list of relevant results is both available and accessible to users, only one result is accessible to Alexa users out of many possible results that may be available. This is somewhat akin to Google's "I'm feeling lucky" feature which takes searchers directly to the first result rather than showing a ranking of results. Prior audits of web search have demonstrated how information diversity and source types can vary when only considering the first or top N results [18, 61]. Another difference between search engines and voice assistants is that web search users typically navigate

results guided by contextual cues such as rankings, brands, timestamps, and other metadata in their determination of what available information to access [61], whereas similar information cues and navigational options are not available for the users of smart assistants. Users are provided only one result that either meets or fails quality expectations in terms of accuracy, relevance, timeliness, completeness, or other informational needs, with little ability to navigate or compare other results. The importance of information quality for a potentially de-contextualised single result is therefore heightened.

The Alexa voice assistant also provides "flash briefings", which are a compilation of news and other content such as comedy, interviews, and fun facts such as "Word of the Day" [17]. In addition to the custom flash briefing provided by Amazon, media outlets and independent developers with the right to distribute original text or audio content can also create and publish their own flash briefings. As of May 2020, there were more than 24,000 news and news-related applications on the Amazon Alexa voice assistant platform, and the numbers continue to rise [2]. Since Amazon provides no gate-keeping, there exist a variety of "unofficial" briefings created by individual developers that provide information from various sources, including those whose credibility may be questionable.

Notwithstanding the prevalence of these flash briefings, this research focuses instead on auditing the more general information response functionality of Alexa. According to the 2019 NPR Smart Audio Report, 74% of 800 surveyed smart speaker owners relied on the device to answer general questions about specific entities [56]. Another study on voice assistants and their implications for news by the Reuters Institute showed that 66% of smart speaker users in the UK used the device to answer general questions compared to 46% of users who mostly used the device to obtain flash briefings [48]. While acknowledging that important future work should address the quality of the wider 3rd party information ecosystem accessible via voice assistants in flash briefings, we scope the current study to only consider responses to news-related queries provided by Alexa's built-in search functionality. This focuses our evaluation on the Alexa voice assistant service itself rather than curated bundles of information supplied by 3rd parties.

4 DATA COLLECTION METHODS

In this section we detail the approach we developed to collect and validate data from voice assistants in response to news-related queries. In particular we detail the development of queries and query phrasings as inputs to the audit, the collection of response data from device, and the assessment of the validity of the data transformation chain in moving from text to audio and back again.

4.1 Generating User Queries

Our user queries are composed of (1) a query topic and (2) a query phrasing. To generate query topics, we fetched the top 20 U.S. trending topics from Google daily search trends for each day of the study. These trends reflect the collective interest and information-seeking behaviour towards topics and events that are capturing people's attention at the current moment, and therefore we expect them to exhibit some degree of newsworthiness [23, 52, 62]. Studies of online news behaviour have shown that web search is a significant origin of users' first article view in a news session [6, 27, 49]. The final topic categories identified this way (see Table 2) exhibit face validity given that they reflect one or more contemporary news values of relevance, including stories about culturally and historically relevant people, organisations, places, and events; the power elite such as powerful individuals, organisations, institutions, or corporations; celebrities and famous people; drama such as scandals, disasters, rescues, or court cases; and entertainment topics encompassing sports, music, movies, comedy and lighter human interest [23].

Following a number of studies that show semantic and syntactic differences between voice and text queries [13, 21], we conducted a crowdsourcing task to better understand potential users' voice

query formulations so as to tailor our audit to common ways in which people seek news-related information from voice assistants. We used Amazon Mechanical Turk (AMT) to crowdsource query phrasings for 144 unique query topics collected from Google daily search trends over a one week period from August 2 to August 9, 2019. For each Human Intelligence Task (HIT), we asked crowd workers to provide three different ways they would ask Alexa for information about each query topic. This approach provides more ecological validity to our audit findings since it aligns the sample of inputs to audit more closely with real user query variation. For each query topic, we requested 5 different workers to complete the task, thereby collecting 15 unique query phrasings for each query topic (as there were three phrasings requested per assignment). Workers were compensated $0.12 for each assignment completed. For all 144 unique query topics, we collected a total of 2,160 query phrasings from 111 distinct workers located in the United States and with assignment approval rate greater than 98% to ensure data reliability. Subject to the population of AMT, which generally reports more "Internet-related experiences, activities, and expertise" than the general population [24], our approach is meant to capture how a relatively broad range of individuals might query the Alexa. Nonetheless, an interesting area for future work would be to examine how query strategies for voice devices may vary demographically, according to skill with the device, or along other attributes, as prior work has begun to do for audits of web search [64].

From the query phrasings collected, we used standard n-gram and word tree visualisation [66] methods to identify frequently occurring query phrasings indicating a general need for information. While the majority of phrasings were worded around specific requests for information, such as "Is Tommy Lee currently married?", and there are many possible query phrases that people can use to ask for information about a specific topic, it would be prohibitive in terms of resources to audit them all. We therefore opted to identify and audit phrasings containing verbs that were general enough to be used with any query topic. As our results will show, these generic query phrasings already suffice to demonstrate substantive variation in resulting responses. The four most-commonly used phrasings we identified include:

- What happened {during / to / in / on / with / at the } _____ ?
- What is going on {during / with / at / in / on} _____ ?
- Can you tell me about {the} _____ ?
- What is new {with / on} _____ ?

After we established the core query phrasings we still needed to select the phrasing with the correct proposition for the specific query topics audited. During the period of the audit we once again obtained query topics from Google trends. Each day at 5:00pm CST, we collected the top 20 daily trending topics from Google and manually prepared the question phrasings by combining the query topic and query phrasing with the appropriate preposition. This step was necessary because we observed that Alexa is less likely to understand queries that do not follow proper preposition use. In a small pilot test we conducted, we observed that removing prepositions from a query such as "What is going on in Indiana?" so the query becomes "What is going on Indiana?" often resulted in the device not able to understand the user query. In most cases, the response was either, "Sorry, I do not understand that" or "I'm not sure". Moreover, it was important to match the type of query topic to a specific preposition (e.g. One might say "What happened *to* {person}" but "What happened *during* {event}). In order to ensure the highest possible validity for the audit we opted to select the preposition manually so we could be sure of its correctness.

4.2 Automated Data Collection

We conducted our study over a two week period from October 28 to November 11, 2019. For each day of our study, we queried the 20 daily trending topics from Google for that day, between 6:00pm

and 9:00pm CST, using all four query phrasings above. For each query topic, we generated four user queries based on the query phrasings above. For example, if one of the Google daily trends was "AOC" (i.e. the initials of U.S. Congresswoman Alexandria Ocasio-Cortez), the four question phrases for that topic would be: (1) "Alexa, what happened to AOC?" (2) "Alexa, what is going on with AOC?" (3) "Alexa, can you tell me about AOC?" and (4) "Alexa, what is new with AOC?" While the Alexa voice assistant is believed to learn from user interactions [5], for example, say one were to ask "Alexa, can you tell me about AOC?" and the voice assistant fails to understand the query or gives an irrelevant response, if a user followed up by saying "Alexa, can you tell me about Alexandria Ocasio-Cortez?", Alexa might learn from this implicit feedback that AOC and Alexandria Ocasio-Cortez are the same entity. To minimise the potential for Alexa to learn from previous queries, we let each response (relevant or not) play uninterrupted, randomised the query order such that the voice assistant cannot tell whether the "user" was satisfied with the response provided or not, and introduced a one minute delay in between each subsequent query. The resulting data set consists of 1, 112 queries and responses. To automate the data collection process, we used the Amazon Web Services (AWS) technology stack. We used Amazon Polly, with default settings and the "Matthew" voice, to synthesise the text of the user queries to speech. Because there is no API for Alexa, we then used the synthesised speech to query a physical smart speaker device (Amazon Echo Plus) running the Alexa voice assistant. From the audio responses, we then used Amazon Transcribe to transcribe the audio responses from the smart speaker. We recorded the text of both the queries and transcribed responses in a database for later analysis[1].

Throughout the study, we considered several technical aspects of Amazon Alexa as well as user factors that might influence results. First, the device was left in its factory default settings for the audit to avoid potential user personalisation effects. Additionally, we used a quiet room with no background noise. For instance, no other people or auditory devices were present in the room at the time of data collection. The same room and microphone and speaker volume settings were also used throughout the study. We further rely on synthesised speech to maintain a consistent acoustic model and control for differences in phonological, phonetic, and prosodic characteristics that have been demonstrated to affect the quality of automated speech recognition [30, 38]. For example, Koenecke's et al [30] study of five state-of-the-art voice assistants (developed by Amazon, Apple, Google, IBM, and Microsoft) demonstrated that all five systems exhibited substantial racial disparities. A separate empirical analysis of bias in voice-based personal assistants [38] also showed interaction bias in users of languages and accents from different geographic regions, particularly those from developing countries. These studies suggest that the quality of interaction via audio depends on many user factors such as language, tone, and accents. In this study we control for these confounding influences by relying on a consistent acoustic model provided by Amazon Polly speech synthesis. Since different smart speaker devices are based on different speech synthesis and recognition technologies by their manufacturers, we align the technology for synthesis and transcription with that of the manufacturer of the smart speaker to control for performance issues in a realistic way. So, just as we use Amazon Polly and Transcribe to audit Alexa here, an audit of Google Home might be more valid if it used Google's technology stack for speech synthesis and transcription.

Finally, there are some limitations worth noting as they suggest important areas for further development. As an initial method focused on adapting to the unique capabilities and the audio modality of voice assistants, we do not take into account other potential confounds such as the location of the device (i.e. possible localisation of information), the configuration of 3rd party

[1]To facilitate replicability of these data collection methods we have made the code available at: https://github.com/comp-journalism/CSCW21-Alexa

information sources, potential memory and learning effects based on previous interactions, or other variations in responses that might be due to factors such as A/B testing. While we rely on the device's factory settings to minimise these effects, methods developed in other audits of information intermediaries have variously dealt with and in some cases measured these confounds [4, 29, 35, 63], and future work should seek to further apply variations in audit methodology to voice assistants.

4.3 Speech Synthesis and Transcription Accuracy

In order to assess the limitations and weaknesses of our automated synthesis and transcription procedure we conducted an accuracy evaluation of the method. From the same 144 query topics that we used to crowd source query phrasings, we synthesised the text of the query topics to speech using Amazon Polly then immediately transcribed the synthesised speech to text using Amazon Transcribe. In both the speech synthesis and transcription processes, the query topics were presented in isolation rather than within a sentence thereby imposing a stringent evaluation by circumventing any potential language models that might infer the correct speech synthesis or word transcription from the context of other words in the sentence. To assess accuracy we compared the transcribed text to that of the original query topic. Using a verbatim query topic match, we found that 77.1% of the query topics were correctly transcribed. Inspecting the responses further, we found that 75% of the incorrectly transcribed query topics appeared to be the result of slang and nicknames rhyming with words, such as Yung Miami ("Young Miami", born Caresha Romeka Brownlee), Yeezy ("Easy", born Kanye Omari West), Bugha ("Bugger", born Kyle Giersdorf), Lori Harvey ("Laurie Harvey"), and Dustin May ("Just in May").

In order to further investigate the nature of errors, we conducted an AMT survey. For each task in this survey, we played audio clips of the voice-synthesised query topics to crowd workers and asked them to classify the pronunciation accuracy of the voice-synthesised text on an ordinal scale of 1 to 3 (1=Completely Incorrect; 2=Mostly Correct; 3=Completely Correct). On a scale of 1 to 5 (least to most difficult), we asked the crowd workers to rate how difficult the query topic was to pronounce. Also on a scale of 1 to 5 (least to most confident), we further asked the crowd workers to rank how confident they were in their classification response and how confident they were that they could correctly pronounce the query topic themselves. For each query topic, we requested 5 different workers to complete the task. Workers were compensated $0.12 for each assignment completed. For all 144 unique query topics, we collected a total of 2,880 responses (as there were four questions per assignment) from 155 distinct workers located in the United States and with an HIT approval rate greater than 98% to ensure data reliability.

In terms of *pronunciation accuracy* we found that correct transcriptions were rated as more accurate ($\mu = 2.85, \sigma = 0.378$) than incorrect transcriptions ($\mu = 2.63, \sigma = 0.614$; $t(250) = 4.65, p < 0.001$), but that both were rated by crowd workers as being mostly to entirely accurately pronounced. Likewise for *pronunciation difficulty* we found that correct transcriptions had lower difficulty ($\mu = 1.44, \sigma = 0.755$), than incorrect ($\mu = 2.01, \sigma = 1.19$) transcriptions ($t(254) = -6.14, p < 0.001$), but that regardless of transcription accuracy the ratings skewed towards the queries being not very difficult to pronounce. This is further mirrored in the confidence assessments of the raters. Correct transcriptions had higher *classification confidence* ($\mu = 4.87, \sigma = 0.370$), than incorrect ($\mu = 4.52, \sigma = 0.819$) transcriptions ($t(224) = 5.85, p < 0.001$), but in both cases confidence is very high. The same observation is true for the crowd workers' confidence in their own ability to correctly pronounce the query topics. While we observed that correct ($\mu = 4.81, \sigma = 0.477$) transcriptions had higher *pronunciation confidence*, than incorrect ($\mu = 4.49, \sigma = 0.825$) transcriptions ($t(244) = 5.02, p < 0.001$), the crowd workers' average confidence remains high in both outcomes. These results suggest that, although the incorrectly transcribed queries tend to be slightly more difficult

to pronounce, they are largely still correctly pronounced. In other words, the errors we observe appear to be mostly driven by transcription issues rather than pronunciation issues.

The implications of these findings are two-fold. First, since some incorrectly understood queries could lead to irrelevant responses, these results suggest that it is important to consider transcription accuracy in audits of smart speakers. These findings help delineate the boundaries of our audit framework in terms of what is technically feasible to evaluate, i.e. the instances of queries in which the voice utterances are correctly pronounced and therefore understood by the system. We therefore incorporate these insights into our analytic framework (described next) by segmenting results according to their transcription quality. Secondly, results on the pronunciation accuracy and difficulty of queries suggest that crowdsourcing utterances of the queries would not lead to an improvement over the method of automated voice-synthesis and transcription we have employed. Indeed, such an approach could be a source of additional confounds as variations in the crowd's pronunciations of query topics (e.g. through differences in accents, volume, pitch, background noises, proximity to the smart speaker, etc) could affect the system's ability to understand the queries, thereby introducing other sources of uncontrolled variance and bias that are otherwise controlled for by our automated method.

5 ANALYTIC FRAMEWORK

We developed an analytic framework for evaluating the data collected during our audit. This framework incorporates the notion of *response rate* as well as dimensions of information quality including *relevance*, *accuracy*, and *timeliness*. We also examine the *source provenance* of information in the results. These dimensions are evaluated according to different *query phrasings* (see Section 4.1) and across different *query categories* (i.e. topics and hard vs. soft news).

In analysing the data collected during our audit we first consider the *response rate*, which is measured as the number of queries that generated a response divided by the total number of queries issued. A non-response indicates a failure to acknowledge the query. This measure is important because it quantifies the extent to which a voice assistant is able to detect and respond to voice commands issued by a user thereby indicating the usability of the device. Of the responses provided, we begin by evaluating whether the smart speaker understood the issued query via a transcription test similar to the one described in section 4.3. Specifically, we measure the transcription quality of the queries issued in the audit and identify correctly transcribed query topics as those "understood" by the voice assistant. Additionally, we consider smart speaker responses in which the query topic is re-iterated as evidence that an utterance has been understood.

After taking into account whether a query was understood by the system, we next develop the framework that we used to evaluate information quality (IQ). In particular we draw on prior work that has defined information quality along several dimensions including accuracy, relevance, timeliness, completeness, format, compatibility, coherence, accessibility, security, and validity [45]. In this work we focus on the relevance, accuracy, and timeliness dimensions as the most crucial for auditing with respect to news-related queries on smart speakers, while deeming the others to be somewhat less pertinent. In applying these in our audit, the first author manually evaluated these dimensions based on online research, after which evaluations were discussed between authors until agreement was reached.

We consider the *response relevance* to be a key component that determines whether the information meets a user's need. If the voice response is irrelevant, users will find the information inadequate regardless of how accurate or timely the information may be. We identify relevant responses as those that relate to the entity in the query topic e.g. when the query response re-iterates the named entity in the query phrasing. We then evaluate information accuracy and timeliness only on the relevant responses. We evaluate the *accuracy* of a response as the extent to which

Relevance	**Definition**: Relevance refers to whether the query response satisfies the user's information need. **Example Relevant**: *Query*: Can you tell me about BlizzCon? *Response*: BlizzCon is an annual gaming convention held by Blizzard Entertainment to promote its major franchises Warcraft, Starcraft, Diablo Hearthstone, Heroes of the Storm and Overwatch. **Example Irrelevant**: *Query*: Can you tell me about the Raiders? *Response*: Paul Revere and the Raiders was an American rock band that saw considerable US mainstream success in the second half of the 1960s and early 1970s.
Accuracy	**Definition**: The extent to which a query response represents the underlying reality about the query topic at any current or prior moment in time. **Example Accurate**: *Query*: What happened to Lil Peep? *Response*: Little Peep, died in Tucson, Arizona, on November 15th. 2017 from drug overdose. He was 21 years old when he died. **Example Inaccurate**: *Query*: What happened at the NCAA? *Response*: There weren't any National Collegiate Athletic Association games during that time.
Timeliness	**Definition**: The degree to which data represent reality from the current point in time i.e. whether the information provided is still current and there is no new information to replace it. **Example Timely**: *Query*: What is new with the impeachment? *Response*: Two U S State Department officials will testify before House of Representatives investigators in the impeachment inquiry against President Donald Trump on Wednesday, while lawmakers prepare to weigh rules formalising the impeachment process ahead. **Example Untimely**: *Query*: What is going on with the impeachment? *Response*: Here's something I found from the article. Impeachment in the United States on Wikipedia. While the actual impeachment of a federal public official is a rare event, demands for impeachment, especially of presidents, are common going back to the administration of George Washington in the mid 17 nineties.

Table 1. Information quality dimensions of relevance, accuracy, and timeliness including definitions and query examples from observed responses.

the information provided reflects the underlying reality of the query topic at any current or prior moment in time. We achieve this by manually looking up the query topic on Google Search and verifying the voice query response with information from a variety of sources including sports, organisations, and government websites as well as news stories and Wikipedia. Only when the voice responses were completely accurate were they labelled as accurate. Finally, we evaluate the *timeliness* of a response based on the degree to which the information provided represents reality from the current point in time i.e. whether the information provided is still current and there is no new information to replace it. Together, these measures reflect information *validity* i.e. whether information is true and satisfies the appropriate standards of relevance, accuracy, and timeliness defined above. Definitions of these measures as well as positive and negative examples to help illustrate them are shown in Table 1.

Our framework also investigates how different *query phrasings* impact the information quality dimensions described above. We therefore evaluate whether the information quality of the query responses varies depending on how the query is phrased. As per the four phrasings identified in section 4.1, we proceed by investigating whether the same query topic, asked differently, affects the relevance, accuracy, and timeliness of the responses. This is important as differences in the

Total Queries (n = 1112)												
People (37.8%)		Sports (19.8)%		Organisations (16.7%)		Entertainment (11.5%)		Events (10.1%)		Products (2.7%)		Locations (1.4%)	
Celebrity	44.8%	Football	67.3%	Sports	80.4%	Movies	53.1%	Holidays	50%	Technology	71.4%	America	50%
Athlete	30.5%	Soccer	16.4%	Business	15.2%	Games	12.5%	Politics	26.9%	Beauty	14.3%	Germany	25%
Politician	16.2%	Basketball	5.5%	Music	2.2%	Music	12.5%	Disasters	7.7%	Information	14.3%	Mexico	25%
Business	3.8%	Other	5.5%	News	2.2%	Comics	9.4%	Entertainment	7.7%				
Journalist	2.9%	Boxing	3.6%			TV	9.4%	Other	3.8%				
Other	1.9%	Cricket	1.8%			Sports	3.1%						

Table 2. Summary statistics of entity categories for query topics. The topics covered a wide variety of issues related to prominent celebrities, athletes, and politicians; holiday and political events; geographic locations; sport and business organisations; as well as technology products, entertainment activities, and other categories (1.1%).

phrasings that people use to seek information from voice assistants have implications in terms of how different demographics may access information [38].

Along the same information quality dimensions described above, we further investigate the extent to which IQ varies across query *categories*. Our query and response data covered seven entity categories (Table 2). These categories include a variety of topics and issues ranging from sports and entertainment to business and politics and coincided with popular holidays such as Diwali and Halloween as well as prominent events such as the 2019 Rugby World Cup and U.S. impeachment probe. The wide range of topics enable us to consider our evaluations within and across entity categories. Furthermore, we evaluate whether there exist information quality differences between hard and soft news categories of queries. We rely on Reinemann's et al [55] definition of *hard news* as politically relevant information with broad societal impacts and *soft news* as information related to the personal lives of celebrities, sports, or entertainment that have no widespread political impact. This distinction is important because it helps to isolate voice assistants' capabilities to provide relevant information about hard news, such as breaking events involving prominent leaders or major social and political issues that have implications for people's ability to understand and engage with civic life. Based on the above definition, we manually identify hard news as query topics that are politically relevant and have broad social impact. These query topics mostly fall within sub-categories such as business, journalist, disasters, politics/politician shown in Table 2. In contrast to hard news, we manually identify soft news as query topics that have no widespread societal impact and mostly fall within sub-categories such as celebrity, sports, games, comics, etc.

Finally, our framework investigates the *provenance* of the information provided by voice assistants as this might indicate both the credibility of the sources as well as the reliability of the information. We thus evaluate whether the same query topic, asked differently, results in responses from different information sources (i.e. provenance), which could indicate different source credibility and information reliability depending on how users phrase their questions. For each question phrasing, we further evaluate the source diversity to determine the extent to which subtle differences in how users interact with voice assistants affect their exposure to diverse information sources.

6 RESULTS

Of the 1,112 queries issued to the Alexa voice assistant, we observe a 92.1% response rate. Of the 88 queries that had no responses, 33 of them were a result of transcription errors. Among the queries with responses we found that 75.9% were "understood" by the device, implying that it could correctly transcribe the query topic. We further observe that on average 71.6% of the understood responses were relevant whereas 16.0% of the understood responses were irrelevant to the questions asked. For the remaining 12.4% of the understood responses, the voice assistant could not provide any useful or informative response and often provided apologetic responses such as "Sorry, I'm not sure about that". Note that "no information" responses are different from a "no response" outcome in which the speaker fails to acknowledge and therefore provide any response to a user query.

We observed that 93.9% of the relevant responses were timely (i.e. up-to-date and representing the present reality of the query topic) and 99.4% of the relevant responses were also accurate in representing the underlying reality about the query topic. While some irrelevant responses were certainly due to transcription errors, even among the understood responses, we observed that irrelevant responses occur when Alexa fails to establish the news context of the query topic. For instance, Alexa responded to the question "What happened at Popeyes?" about a man killed over a chicken sandwich at a Popeyes fast-food restaurant [8] in Maryland, USA, with an answer about the cartoon character Popeye. This example, among other similar instances, illustrates a gap in functionality where the voice assistant is unable to provide relevant information that reflects the underlying reality about an entity from the current point in time, perhaps reflecting a lack of understanding of the user's query intent [31] or shared common ground [10, 12, 46] on the query context. We further explore this issue in the discussion.

In the following sections, we show how the information quality and sources of Alexa's responses vary depending on the way that a query is phrased thereby demonstrating how subtle variations in query phrasings affect the smart speaker's ability to understand users' queries and provide relevant and timely responses. We further show how the sources of the responses provided by Alexa also vary by query phrasing and identify a significant shortcoming of the lack of provenance information in responses provided by Alexa. The current findings are limited as they only apply to one voice assistant, Amazon Alexa. Voice assistants from other providers may exhibit better or worse performance in terms of information quality hence future work might be informed by the methods developed here in running a comparative study across many devices.

6.1 Does information quality vary by query phrasing?

In our study, we observe that the question phrasings "Can you tell me about" and "What happened" had higher proportions of relevant responses compared to "What is going on" and "What is new" (See Figure 2). This indicates that Alexa appears to perform better in finding relevant information that is static, and struggles to find relevant information about new and evolving situations currently in the news. For example, when asked "Can you tell me about" or "What happened to" Jeff Sessions, Alexa responded with information from the U.S. politician's Wikipedia article. However, when asked "What is going on with" or "What is new with" Jeff Sessions, Alexa could not find any information about the politician. Given the news context of Jeff Sessions at the time of data collection (a former U.S. Attorney General who was fired), even general phrasings like "Can you tell me about" and "What happened to" might warrant a news-related, rather than encyclopedic response. And query phrasings such as "What is going on with" and "What is new with" more directly allude to an information request that is related to the present moment. Additionally, when evaluating the impact of different query phrasings on information timeliness, we observe a 6.2% decrease from response relevance to response timeliness for the phrasing "What happened" compared to a 3.7% and 3.5% decrease for the phrases "What is going on" and "What is new", respectively (Table 3).

6.2 Does information quality vary by news category?

We further investigated whether the information quality of news-related queries varies by news category. An analysis of the response relevance by the query topic categories (Table 4) reveals that sports topics have the highest response relevance (87.0%), followed by organisations (84.7%). However, it is important to highlight that 80.4% of the organisations were in the sports sub-category (Table 2) e.g. the U.S. National Football League (NFL) or Major League Baseball (MLS). Sports topics also had the highest response timeliness (84.6%). A closer examination of our data shows that when it comes to sports topics for example, the Alexa voice assistant can provide reliable and up-to-the-minute updates, but often fails to provide timely information about other events (47.9%),

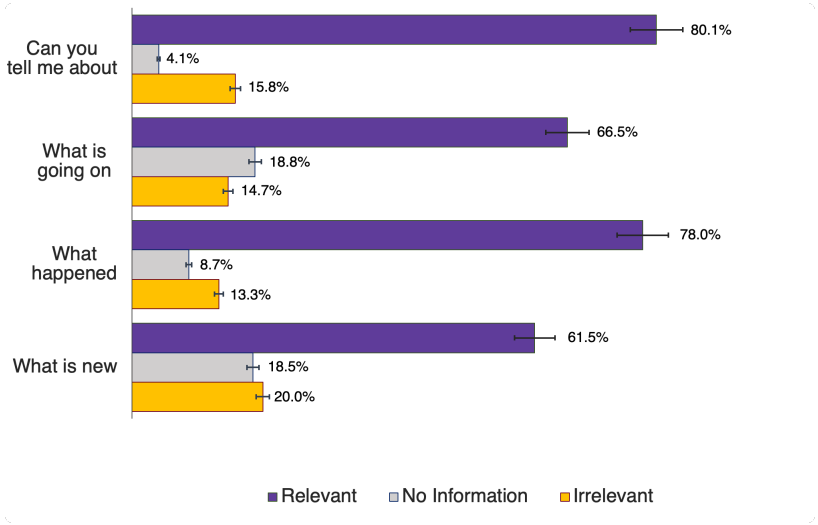

Fig. 2. Response relevance (%) of all understood responses broken down by question phrasing. The same question, phrased differently, yields different rates of response relevance.

Query Phrasing	Response Rate (Count)	Understood Rate (Count)	Relevance Rate (Count)	Accuracy Rate (Count)	Timeliness Rate (Count)
Can you tell me about	93.5% (260)	75.4% (196)	80.1% (157)	79.1% (155)	76.0% (149)
What is new	93.5% (260)	75.0% (195)	61.5% (120)	61.0% (119)	57.9% (113)
What happened	91.7% (255)	76.5% (195)	78.0% (152)	78.0% (152)	71.8% (140)
What is going on	89.6% (249)	76.7% (191)	66.5% (127)	66.5% (127)	62.8% (120)
All phrasings	92.1% (1024)	75.9% (777)	71.6% (556)	71.2% (553)	67.2% (522)

Table 3. Response, understood, relevance, accuracy, and timeliness rates and counts for each query phrasing. Note that the understood rate is based on response count, and the relevance, accuracy, and timeliness rates are based on the understood count.

especially those that are politically relevant. For example, when asked "What is going on with Panthers vs Packers" the Alexa voice assistant responded, "Currently, the Packers are beating the Panthers, 24 - 16 with 19 seconds left in the fourth quarter" and when asked "What is going on with the Impeachment", the voice assistant responded, "Here's something I found from the article 'Impeachment in the United States on Wikipedia': While the actual impeachment of a federal public official is a rare event, demands for impeachment, especially of presidents, are common going back to the administration of George Washington in the mid 1790s."

While the query topic "the impeachment" might refer to any number of impeachments or even simply the noun itself, since we use query phrasings that indicate interest in current information (e.g. "What is *going on* with the impeachment"), it is reasonable to anticipate a news-related response about the U.S. impeachment trial of Donald Trump from the voice assistant. However, despite providing contextual cues to elicit news-related responses a communication breakdown occurs when the voice assistant lacks enough evidence to coordinate its knowledge state with that of the user. This issue is particularly surprising for the products category, which had the lowest response relevance (37.5%), since it is mostly comprised of technology-related products such as the AppleTV or AirPods, whose information likely already readily exists in structured databases available to Amazon.

Category	Response Rate (Count)	Understood Rate (Count)	Relevance Rate (Count)	Accuracy Rate (Count)	Timeliness Rate (Count)
People	91.0% (382)	70.9% (271)	69.0% (187)	69.0% (187)	65.3% (177)
Sports	91.4% (201)	80.6% (162)	87.0% (141)	87.0% (141)	84.6% (137)
Organisations	93.6% (174)	75.3% (131)	84.7% (111)	84.0% (110)	83.2% (109)
Entertainment	94.5% (121)	82.6% (100)	54.0% (54)	53.0% (53)	51.0% (51)
Events	91.1% (102)	71.6% (73)	61.6% (45)	60.3% (44)	47.9% (35)
Products	93.3% (28)	85.7% (24)	37.5% (9)	37.5% (9)	33.3% (8)
Locations	100% (16)	100% (16)	56.3% (9)	56.3% (9)	31.3% (5)
All categories	92.1% (1024)	75.9% (777)	71.6% (556)	71.2% (553)	67.2% (522)

Table 4. Response, understood, relevance, accuracy, and timeliness rates and counts for each query topic category. Note that the understood rate is based on response count, and the relevance, accuracy, and timeliness rates are based on the understood count.

We also observed significant differences in information quality associated with the amount of politically-relevant query topics in each news category. For example, in the people and events categories that include a politics sub-category, we observe that the higher the proportion of politically-relevant query topics (16.2% and 26.9% respectively), the lower the response relevance (69.0% and 61.6% respectively) for the topic ($\chi^2(1, N = 324) = 196.9, p < 0.01$). Whereas the low response relevance of events-related query topics emphasises challenges in the smart speaker's ability to support query topics that are "of the moment", the observed differences in information quality proportion of politically-relevant query topics motivated us to investigate whether there exist information quality differences between hard and soft news. Our results show that hard news and soft news have equal response rates compared to all the responses. However, we notice substantial differences in the response relevance of hard and soft news query topics. Specifically, while the response rates for both hard and soft news are consistent with that of all the responses, hard news had a 41.0% (68 of 166) response relevance whereas soft news had a 56.9% (488 of 858) response relevance. Additionally, hard news had a lower response timeliness 36.8% compared to soft news 53.7%.

6.3 Do response sources vary by query phrasing?

Finally, we investigated whether the same query topic, phrased differently resulted in responses from different information sources (Figure 3). Our results show that, for the most part, query responses lack provenance information as indicated by the large proportion of unknown sources. Overall, we observed that 60.4% of all understood responses were of unknown provenance. The question phrasing "Can you tell me about", provides the least number of unknown sources and most number of Wikipedia sources. We further observe that the phrasing "What is new" provides the most number of news sources. In our response data, all news sources were either Reuters or the Washington Post.

The lack of information provenance for many responses hampers a comprehensive audit of the credibility of information sources. Of all the understood responses, Wikipedia is the most prevalent individual information source, providing 18.6% of the responses. It is plausible to conclude that the reliability of these responses is only as reliable as Wikipedia. While 14.6% of the understood responses were generic Alexa responses about the information that the voice assistant could not find, the remaining 6.4% of the responses were from a variety of sources of varying credibility such as the Washington Post, Reuters, IMDb, World Atlas, Amazon customers, and Reference.com. Among these sources news sources accounted for only 1.4% of the sources in understood responses, including Reuters (1.2%) and The Washington Post (0.2%). Responses from Amazon were either

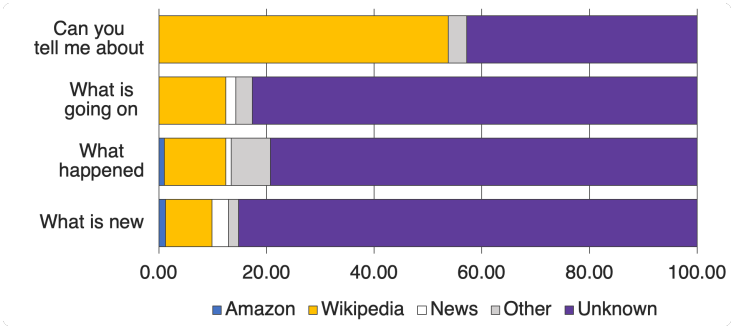

Fig. 3. Source diversity (%) of understood responses broken down by question phrasing. The same question, phrased differently, may prompt a response from a different information source.

references to Amazon music albums, Amazon deals, or crowd sourced responses from Amazon customers. It is possible that the source of information varies based on whether any third-party skills such as Yelp or AccuWeather are enabled. However as noted previously, third-party skills were disabled in our evaluation in order to focus on the built-in information Alexa relies on.

7 DISCUSSION

In this paper, we present a methodological approach for evaluating the information quality of news-related queries on the Amazon Alexa voice assistant. Our methodological approach illustrates the need to tailor audits of voice assistants to the specific capabilities and modalities of these systems, taking into consideration the confounding influences of user factors such as variations in language tone, accents, and pronunciation accuracy that may affect the voice assistant's speech recognition capacity to accurately listen, process, and respond to users' queries. The audit we have presented here, albeit modest in duration and data collection, is sufficient to both demonstrate the applicability of our methodological framework for auditing voice assistants, and to illuminate some of the issues of information quality that the Alexa voice assistant struggles with. In the following subsections we elaborate the implications of the results we observed in our audit of Alexa in terms of what they mean for information quality for different types of information and users, user intent and grounding for voice-based news search, and information source and provenance issues.

7.1 Information Quality on Alexa

We find that Alexa is able to return relevant responses for 71.6% of the understood queries we audited, suggesting moderate performance in being able to respond to news-related topics provided the query topic itself was understood (which itself was only about three-quarters of the total number of queries issued). However, amongst the relevant responses, the majority were of high information quality (93.9% timely and 99.4% accurate), with some substantive variations according to query phrasing and news category. Hard news topics suffer from worse information quality, including relevance and timeliness, compared to soft news topics, suggesting that Alexa users may face additional constraints with respect to finding information about societal and politically relevant events compared to sports, entertainment and lifestyle events.

Moreover, some users may face degraded experiences and difficulties accessing timely and reliable information due to subtle differences in how they phrase queries. While these results are consistent with findings from audits of web search engines that show some variations in results based on query formulation [61, 64], our analysis here is further able to connect such variations

to issues of information quality. For instance the "What is new" query phrasing resulted in the lowest relevance and highest irrelevance scores in comparison to the other phrasings we tested, and "What happened" resulted in the largest drop in timeliness. This suggests that different question phrasings on the same topic could lead to certain users potentially and unknowingly suffering worse user experience and information access compared to others, depending on subtle differences in how they interact with the voice assistant. To overcome this problem, voice assistants may need to be designed and evaluated such that they respond consistently across variations in query phrasings [37, 54].

Similarly, our results from crowd-based human judgements of pronunciation classification, difficulty, and confidence of query topics help demonstrate that information about entities whose names are difficult to pronounce and/or transcribe automatically may not be as easy to obtain. Future research might seek to further elaborate whether there are patterns between question phrasing and other relevant factors such as education level, social-economic standing, or cultural background, which could lead to unequal access to information based on the phrasing patterns of certain user groups. In turn, similar studies of users' needs and information seeking behaviour in voice-based search can inform future audits of information quality in smart speakers taking into account particular conversational styles of different demographics.

7.2 User Intent and Grounding for the News Domain

Some of the voice assistant's challenges to providing relevant, accurate, and timely responses may be due to the way in which the transcribed text is used to search for the single result response provided to users. Although little is known publicly about how the Amazon Alexa smart speaker functions to fulfil users' information needs, previous studies have shown that compared to web-based search, voice search yields worse information retrieval performance if the transcribed queries are issued unmodified to search APIs [3]. While simple pre-processing techniques through parts-of-speech (POS) tagging have been shown to improve the information retrieval process in conversational search [13], we anticipate that these challenges are likely beyond keyword extraction and may require rule-based or machine learning based models for mapping users' utterances to query categories for intent classification. Problems of query intent classification are not unique to voice-assistants. For example, prior work has proposed click-based models for estimating whether a web-based search query is news-related so that a web search engine can specifically include news-related results [31, 51]. But classifying a query's intent is a more challenging task for conversational search because there are no user click-through behaviours to learn from. Future smart speaker development might, however, consider intent-classification mechanisms that leverage complementary data sources such as news aggregators or web search engines.

Another potential way that voice assistants might better establish query intent is through conversational grounding approaches. It is reasonable to believe that many of the errors that occur with the irrelevant and untimely responses that were correctly transcribed can be explained as failures of conversation grounding—the process by which interlocutors establish shared understandings of their respective knowledge and intents [10, 12, 46]. In other words, if Alexa were able to understand the intent of a user in asking for news information it might be better able to respond with relevant and timely information, perhaps from established and credible news organisations. A multi-turn interaction that facilitates conversation grounding between a user and the voice assistant could help to correctly classify the user's query intent e.g. whether a user is searching general or news-related information. This process could help the device coordinate its knowledge states with that of the user as expressed through distinct query phrasings intended to elicit news-related responses. To support better human-computer interactions with voice assistants, one area for future research is to develop and study effective conversational grounding techniques. Rather than aiming to quickly

identify the single most relevant and accurate response to a user's query, voice assistants might benefit from more conversational interactions that facilitate multiple turn-taking to enable response clarification and refinement by suggesting response categories and eliciting feedback [37]. As voice assistants adopt such techniques however, this will in turn demand that auditing methods and protocols adapt to support structured "query conversations" that reflect typical human behavior.

7.3 Information Provenance Issues

In our audit, we found that 60.4% of the query responses we observed lacked provenance information. The predominant absence of provenance information creates challenges for users in terms of how to ascertain the reliability and credibility of information sources. We believe it is important for voice assistants to provide provenance information for each of their query responses so as to signal the evidence on which the information is based. Information provenance is crucial in voice-based search because in addition to enabling users to identify the source of information, users could further improve their evaluation of the credibility of the voice assistant. Similar to web-based search whereby users can interact with graphical interfaces and make use of rich contextual cues to navigate through a set of results and find the most relevant to their information need, voice assistants could benefit from providing responses that not only reveal the provenance of the information but also offer clues about the underlying algorithmic process that produced the query response (e.g. "Here is a news article about [topic] from [source] that [was published a few hours ago / is currently popular online]) [10]. In turn, these contextual cues can help users build mental models of the search space and better understand the voice assistant's capabilities and limitations and ultimately enable users to formulate better queries to elicit more relevant, accurate, and timely information [37]. Provenance information might also be conveyed conversationally, through follow-up questions like "Alexa, how do you know that?"

Furthermore we found that there were very few responses that referred to news sources (1.4% from only two sources), which is somewhat surprising given the nature of the queries we audited. While we cannot tell for certain why Alexa provides responses from such a limited set of news sources, other audits of web-search algorithms have also demonstrated evidence of source concentration whereby a few news outlets are responsible of an outsize majority of the top-ranking search results that are visible to users [27, 28, 63]. One way that smart speakers might be improved in the future is to better incorporate structured knowledge extracted from a variety of news sources in an ongoing, timely way. Better provision of information sourcing would also unlock more possibilities for audits, which could look at source location, type, or political bias [4, 18, 63].

Our findings also showed that that the majority of the responses that did have a known source were from Wikipedia. This is concerning because even though, in some cases, Wikipedia may provide reliable information, the peer-production nature of the open knowledge community means that any information it contains at any given time may be vandalised, incomplete, or inaccurate. Moreover, prominent news subjects as well as politically and culturally contentious topics in Wikipedia are often vulnerable to vandalism during 'edit wars' and although edit errors and conflicts may be resolved, some errors may remain unnoticed for long periods of time, especially considering that Wikipedia is volunteer run [67]. These findings also feed into the ongoing debate about the peculiar relationship between user-generated content and algorithmic information intermediaries [65], specifically the concern over what commercial entities such as Amazon may "owe" to the creators of freely available and volunteer-created information that their technologies depend on.

8 CONCLUSION

This paper presents a data-driven methodological approach for evaluating the information quality of responses to news-related queries on voice assistant devices. The approach we introduce demonstrates how to tailor audits of voice assistants to the system's specific capabilities and input modalities (e.g. choice of input phrasing, automated speech synthesis and transcription, analytic framework involving response rate and understood rate), and presents a framework for evaluating information quality according to relevance, accuracy, timeliness, and source provenance that is sensitive to query phrasing and news query type. By describing the method in detail we hope that future research can adopt, adapt, and apply it to other smart speaker devices.

We demonstrated the applicability of the method we developed through an audit of the Alexa voice assistant. We found that, overall, the majority of Alexa's responses to queries that were understood were relevant, and of those they were predominantly accurate and timely. However, we also observed a fair amount of variation based on query phrasings and query categories, pointing out areas where the information quality of responses is reduced, such as for hard news queries. Moreover, we identify a troubling lack of information provenance for most responses, which challenges users' ability to effectively evaluate the credibility and source of information. These findings present key concerns for information quality on the Alexa voice assistant that merit further investigation. Future studies should aim to investigate these and other concerns on different voice assistants e.g. Google Home, Apple's Siri, Microsoft's Cortana, etc. As more and more people increasingly use smart speakers to seek news-related information, audits of information quality in smart speakers are crucial to investigate potential misinformation risks that have broad societal consequences. Periodic auditing of information quality from voice assistants may be necessary, or could be imagined as a more fully integrated aspect of the engineering development process [54].

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation Grant, Award IIS-1717330. The authors would like to thank Sophie Liu, Victoria Cabales, Benjamin Scharf, and the Knight Lab Studio at Northwestern University for their support and assistance in a related pilot study.

REFERENCES

[1] Amazon. 2017. Creating IoT Solutions With Serverless Architecture & Alexa. Retrieved from https://www.slideshare.net/AmazonWebServices/creating-iot-solutions-with-serverless-architecture-alexa. Accessed: 2020-09-22.

[2] Amazon. 2020. Amazon Alexa Search: "News". Retrieved from https://alexa.amazon.com/spa/index.html#skills/search/news/?&ref-suffix=sb_gw. Accessed: 2020-09-22.

[3] J Arguello, B Choi, and R Capra. 2017. Factors affecting users' information requests. In *SIGIR 1st International Workshop on Conversational Approaches to Information Retrieval (CAIR'17)*, Vol. 4.

[4] Jack Bandy and Nicholas Diakopoulos. 2020. Auditing news curation systems: A case study examining algorithmic and editorial logic in Apple News. *Proc. International Conference on Web and Social Media (ICWSM)* (2020).

[5] Bryan Barrett. 2018. The Year Alexa Grew Up. Retrieved from https://www.wired.com/story/amazon-alexa-2018-machine-learning. Accessed: 2020-09-18.

[6] Frank Bentley, Katie Quehl, Jordan Wirfs-Brock, and Melissa Bica. 2019. Understanding online news behaviors. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.

[7] Engin Bozdag. 2013. Bias in algorithmic filtering and personalization. *Ethics and Information Technology* 15, 3 (2013), 209–227.

[8] Lynh Bui. 2019. Popeyes Stabbing Suspect Still Sought in Killing That May Have Stemmed From Fight Over Popular Chicken Sandwich. Retrieved from https://www.washingtonpost.com/local/public-safety/attacker-still-sought-in-popeyes-killing-that-may-have-stemmed-from-fight-over-popular-chicken-sandwich/2019/11/05/fb2c29e2-ffd6-11e9-9777-5cd51c6fec6f_story.html. Accessed: 2020-08-11.

[9] Daniel Bush and Alex Zaheer. 2019. Bing's Top Search Results Contain an Alarming Amount of Disinformation. Retrieved from https://cyber.fsi.stanford.edu/io/news/bing-search-disinformation. Accessed: 2020-10-08.

[10] Janghee Cho and Emilee Rader. 2020. The role of conversational grounding in supporting symbiosis between people and digital assistants. *Proceedings of the ACM on Human-Computer Interaction* CSCW (2020).
[11] Hyunji Chung, Michaela Iorga, Jeffrey Voas, and Sangjin Lee. 2017. Alexa, can I trust you? *Computer* 50, 9 (2017), 100–104.
[12] Herbert H Clark and Susan E Brennan. 1991. *Grounding in Communication.* American Psychological Association.
[13] Fabio Crestani and Heather Du. 2006. Written versus spoken queries: A qualitative and quantitative comparative analysis. *Journal of the American Society for Information Science and Technology* 57, 7 (2006), 881–890.
[14] Henry K Dambanemuya and Nicholas Diakopoulos. 2020. "Alexa, what is going on with the impeachment?" Evaluating smart speakers for news quality. In *Proceedings of the Computation and Journalism Symposium.* 1–4.
[15] Nicholas Diakopoulos. 2019. *Automating the News: How Algorithms Are Rewriting the Media.* Harvard University Press.
[16] Nicholas Diakopoulos, Daniel Trielli, Jennifer Stark, and Sean Mussenden. 2018. I vote for—how search informs our choice of candidate. *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple, M. Moore and D. Tambini (Eds.)* 22 (2018).
[17] Amazon Developer Documentation. 2020. Understand the Flash Briefing Skill API. Retrieved from https://developer.amazon.com/en-US/docs/alexa/flashbriefing/understand-the-flash-briefing-skill-api.html. Accessed: 2020-09-22.
[18] Sean Fischer, Kokil Jaidka, and Yphtach Lelkes. 2020. Auditing local news presence on Google News. *Nature Human Behaviour* 12 (2020), 1 – 9. https://doi.org/10.1038/s41562-020-00954-0
[19] Nathaniel Fruchter and Ilaria Liccardi. 2018. Consumer attitudes towards privacy and security in home assistants. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems.* ACM.
[20] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media.* Yale University Press, New Haven, CT.
[21] Ido Guy. 2018. The characteristics of voice search: Comparing spoken with typed-in mobile web search queries. *ACM Transactions on Information Systems (TOIS)* 36, 3 (2018), 1–28.
[22] Mario Haim, Andreas Graefe, and Hans-Bernd Brosius. 2017. Burst of the filter bubble? *Digital Journalism* 6, 3 (July 2017), 1–14.
[23] Tony Harcup and Deirdre O'Neill. 2017. What is news? News values revisited (again). *Journalism Studies* 18, 12 (2017), 1470–1488.
[24] Eszter Hargittai and Aaron Shaw. 2020. Comparing internet experiences and prosociality in Amazon Mechanical Turk and population-based survey samples. *Socius: Sociological Research for a Dynamic World* 6 (2020), 237802311988983. https://doi.org/10.1177/2378023119889834
[25] Jiepu Jiang, Ahmed Hassan Awadallah, Rosie Jones, Umut Ozertem, Imed Zitouni, Ranjitha Gurunath Kulkarni, and Omar Zia Khan. 2015. Automatic online evaluation of intelligent assistants. In *Proc. International Conference on World Wide Web (WWW).*
[26] Pascal Juergens and Birgit Stark. 2017. The power of default on Reddit: A general model to measure the influence of information intermediaries. *Policy & Internet* 9, 4 (2017), 395–419.
[27] Anna Kawakami, Khonzoda Umarova, Dongchen Huang, and Eni Mustafaraj. 2020. The fairness doctrine lives on? Theorizing about the algorithmic news curation of Google's top stories. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media.* 59–68.
[28] Anna Kawakami, Khonzodakhon Umarova, and Eni Mustafaraj. 2020. The media coverage of the 2020 US Presidential Election candidates through the lens of Google's top stories. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 868–877.
[29] Chloe Kliman-Silver, Aniko Hannak, David Lazer, Christo Wilson, and Alan Mislove. 2015. Location, location, location: The impact of geolocation on Web search personalization. In *Proceedings of the 2015 Internet Measurement Conference.* 121–127.
[30] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences* (2020).
[31] Arnd Christian König, Michael Gamon, and Qiang Wu. 2009. Click-through prediction for news queries. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval.* 347–354.
[32] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P. Gummadi, and Karrie Karahalios. 2017. Quantifying search bias: Investigating sources of bias for political searches in social media. In *Proc. Conference on Computer Supported Cooperative Work and Social Computing (CSCW).*
[33] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P Gummadi, and Karrie Karahalios. 2019. Search bias quantification: Investigating political bias in social media and web search. *Information Retrieval Journal* 22, 1-2 (2019), 188–227.
[34] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. 2018. Alexa, are you listening?: Privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proceedings of the ACM on Human-Computer Interaction*

2, CSCW (2018), 102.
[35] Huyen Le, Raven Maragh, Brian Ekdale, Andrew High, Timothy Havens, and Zubair Shafiq. 2019. Measuring political personalization of Google News search. In *The World Wide Web Conference*. 2957–2963.
[36] Chang Li and Hideyoshi Yanagisawa. 2019. Intrinsic motivation in virtual assistant interaction. In *International Symposium on Affective Science and Engineering*. Japan Society of Kansei Engineering, 1–5.
[37] Q Vera Liao, Werner Geyer, Michael Muller, and Yasaman Khazaen. 2020. Conversational Interfaces for Information Search. In *Understanding and Improving Information Search*. Springer, 267–287.
[38] Lanna Lima, Vasco Furtado, Elizabeth Furtado, and Virgilio Almeida. 2019. Empirical analysis of bias in voice-based personal assistants. In *Companion Proceedings of the 2019 World Wide Web Conference*. ACM, 533–538.
[39] Irene Lopatovska, Katrina Rink, Ian Knight, Kieran Raines, Kevin Cosenza, Harriet Williams, Perachya Sorsche, David Hirsch, Qi Li, and Adrianna Martinez. 2018. Talk to me: Exploring user interactions with the Amazon Alexa. *Journal of Librarianship and Information Science* (2018).
[40] Gustavo López, Luis Quesada, and Luis A Guerrero. 2017. Alexa vs. Siri vs. Cortana vs. Google Assistant: A comparison of speech-based natural user interfaces. In *International Conference on Applied Human Factors and Ergonomics*. Springer, 241–250.
[41] Silvia B Lovato, Anne Marie Piper, and Ellen A Wartella. 2019. Hey Google, do unicorns exist?: Conversational agents as a path to answers to children's questions. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children*. ACM, 301–313.
[42] Mykola Makhortykh, Aleksandra Urman, and Roberto Ulloa. 2020. How search engines disseminate information about COVID-19 and why they should do better. *Harvard Kennedy School Misinformation Review* 1, 3 (2020). https://doi.org/10.37016/mr-2020-017
[43] Danaë Metaxa, Joon Sung Park, James A Landay, and Jeff Hancock. 2019. Search media and elections: A longitudinal investigation of political search results. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–17.
[44] P T Metaxas and Y Pruksachatkun. 2017. Manipulation of search engine results during the 2016 US Congressional Elections. *Proc International Conference on Internet and Web Applications (ICIW)* (2017).
[45] Holmes Miller. 1996. The multiple dimensions of information quality. *Information Systems Management* 13, 2 (1996), 79–82.
[46] Andrew Monk. 2003. Common ground in electronically mediated communication: Clark's theory of language use. *HCI Models, Theories, and Frameworks: Toward a Multidisciplinary Science* (2003), 265–289.
[47] Philip Napoli. 2019. *Social Media and the Public Interest: Media Regulation in the Disinformation Age*. Columbia University Press, New York, NY.
[48] Nic Newman. 2018. *The Future of Voice and the Implications for News*. Reuters Institute for the Study of Journalism.
[49] Nic Newman, Richard Fletcher, Anna Schulz, Simge Andi, and Rasmusen Nielsen. 2020. *Reuters Institute Digital News Report*. Reuters Institute for the Study of Journalism.
[50] Dimitar Nikolov, Mounia Lalmas, Alessandro Flammini, and Filippo Menczer. 2019. Quantifying biases in online information exposure. *Journal of the Association for Information Science and Technology* 70, 3 (2019), 218–229.
[51] Rajesh Parekh, Jignashu Parikh, and Pavel Berkhin. 2009. Predicting Newsworthy Queries Using Combined Online and Offline Models. US Patent App. 12/104,111.
[52] Poynter. 2015. The new Google Trends is a real-time news detection system. Library Catalog: www.poynter.org Section: Newsletters.
[53] Cornelius Puschmann. 2018. Beyond the bubble: Assessing the diversity of political search results. *Digital Journalism* 10, 24 (Nov. 2018), 1–20.
[54] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery.
[55] Carsten Reinemann, James Stanyer, Sebastian Scherr, and Guido Legnante. 2012. Hard and soft news: A review of concepts, operationalizations and key findings. *Journalism* 13, 2 (2012), 221–239.
[56] NPR & Edison Research. 2019. The Smart Audio Report.
[57] NPR & Edison Research. 2020. The Smart Audio Report.
[58] Ronald E Robertson, David Lazer, and Christo Wilson. 2018. Auditing the personalization and composition of politically-related search engine results pages. In *Proc. World Wide Web Conference (WWW)*.
[59] Jan-Hinrik Schmidt, Lisa Merten, Uwe Hasebrink, Isabelle Petrich, and Amelie Rolfs. 2019. How do intermediaries shape news-related media repertoires and practices? Findings from a qualitative study. *International Journal of Communication* 13 (2019), 21.
[60] Alex Sciuto, Arnita Saini, Jodi Forlizzi, and Jason I Hong. 2018. Hey Alexa, what's up?: A mixed-methods studies of in-home conversational agent usage. In *Proceedings of the 2018 Designing Interactive Systems Conference*. ACM,

857–868.
[61] Miriam Steiner, Melanie Magin, Birgit Stark, and Stefan Geiß. 2020. Seek and you shall find? A content analysis on the diversity of five search engines' results on political queries. *Information, Communication & Society* (2020), 1–25.
[62] Filippo Trevisan, Andrew Hoskins, Sarah Oates, and Dounia Mahlouly. 2018. The Google voter: Search engines and elections in the new media ecology. *Information, Communication & Society* 21, 1 (2018), 111–128.
[63] Daniel Trielli and Nicholas Diakopoulos. 2019. Search as news curator: The role of Google in shaping attention to news information. In *Proc. Conference on Human Factors in Computing Systems (CHI)*.
[64] Daniel Trielli and Nicholas Diakopoulos. 2020. Partisan search behavior and Google results in the 2018 US midterm elections. *Information, Communication & Society* (2020), 1–17.
[65] Nicholas Vincent, Isaac Johnson, Patrick Sheehan, and Brent Hecht. 2019. Measuring the importance of user-generated content to search engines. In *Proc. International Conference on Web and Social Media (ICWSM)*.
[66] Martin Wattenberg and Fernanda B Viégas. 2008. The word tree, an interactive visual concordance. *IEEE Transactions on Visualization and Computer Graphics* 14, 6 (2008), 1221–1228.
[67] Wikipedia. 2019. Wikipedia Is Not a Reliable Source. Retrieved from https://en.wikipedia.org/wiki/Wikipedia:Wikipedia_is_not_a_reliable_source. Accessed: 2021-01-12.

Received June 2020; revised October 2020; accepted December 2020