Misinformation in Third-party Voice Applications

Mary Bispham King's College London London, UK

Xavier Ferrer-Aran King's College London London, UK Suliman Kalim Sattar King's College London London, UK

Jide Edu University of strathclyde Glasgow, UK

Jose Such King's College London London, UK

1 INTRODUCTION

ABSTRACT

This paper investigates the potential for spreading misinformation via third-party voice applications in voice assistant ecosystems such as Amazon Alexa and Google Assistant. Our work fills a gap in prior work on privacy issues associated with third-party voice applications, looking at security issues related to outputs from such applications rather than compromises to privacy from user inputs. We define misinformation in the context of third-party voice applications and implement an infrastructure for testing third-party voice applications using automated natural language interaction. Using our infrastructure, we identify - for the first time - several instances of misinformation in third-party voice applications currently available on the Google Assistant and Amazon Alexa platforms. We then discuss the implications of our work for developing measures to pre-empt the threat of misinformation and other types of harmful content in third-party voice assistants becoming more significant in the future.

CCS CONCEPTS

• Security and privacy \rightarrow Human and societal aspects of security and privacy; • Human-centered computing \rightarrow Human computer interaction (HCI).

KEYWORDS

voice assistants, online harm, misinformation

ACM Reference Format:

Mary Bispham, Suliman Kalim Sattar, Clara Zard, Xavier Ferrer-Aran, Jide Edu, Guillermo Suarez-Tangil, and Jose Such. 2023. Misinformation in Third-party Voice Applications. In ACM conference on Conversational User Interfaces (CUI '23), July 19–21, 2023, Eindhoven, Netherlands. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3571884.3604307

CUI '23, July 19–21, 2023, Eindhoven, Netherlands

© 2023 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-0014-9/23/07.

https://doi.org/10.1145/3571884.3604307

The increasing use of voice assistants and other voice-controlled devices has brought about a number of new security concerns. A substantial amount of research has been conducted on the vulnerability of voice interfaces to attacks and the threat to privacy from voice assistants — see [19] for a recent literature review. Some of this work focuses specifically on security concerns associated with voice applications published by third-party developers [10, 16, 39]. Users access third-party voice applications via the voice interface of a core assistant such as Google Assistant or Amazon Alexa. Third-party voice applications in the Google Assistant ecosystem are known as Actions, whereas third-party voice applications in the Alexa ecosystem are known as Skills. Actions and Skills are available in various categories for purposes such as education, fitness, games, and others [18].

As third-party voice applications grow in use and sophistication, their security risks also increase [19]. Processes for assessing and enforcing the security of third-party voice applications are still in the early stages of development. Security issues associated with third-party voice applications may arise with regard to user input requested by an application, as well as with regard to harmful outputs from an application. An example of the former type of issue is illegitimate requests for personal information via the conversational interface of an application [17, 21, 27, 34]. In contrast, an example of the latter is the spread of misinformation via the conversational interface, as we study in this paper. To the best of our knowledge, no instances of misinformation in third-party voice applications have been recorded at the time of writing.

In this paper, we propose a method for measuring misinformation via third-party voice applications and finding instances of such content in Google Actions and Alexa Skills in the wild. Given the prevalence of misinformation and other harmful content in other digital ecosystems such as social media platforms [8, 41], it is reasonable to expect that such content may also be embedded in voice assistant ecosystems via third-party voice applications.

2 THREAT MODEL

Voice assistant providers typically vet third-party voice applications to ensure they do not contain inappropriate content before publication. Notwithstanding, providers like Amazon and Google

Clara Zard King's College London London, UK

Guillermo Suarez-Tangil IMDEA Networks Institute Madrid, Spain

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

do not reveal how they vet these applications. However, it is apparent that testing of third-party applications' back-end code running on developers' servers is limited to dynamic black-box testing, as providers do not have direct access to this code. Thus, the security of these vetting processes are inadequate [2, 10, 17, 21, 43]. Aside from issues around the effectiveness of security vetting processes for third-party voice applications, developers can also modify the back-end code of the voice applications to implement malicious functionality after such applications have been vetted and published [10, 27, 39].

Misinformation is particularly problematic in third-party voice applications because users seem to have higher trust in assistants due to their anthropomorphic characteristics [12, 36] and consider them reliable in terms of information than a regular web search [14, 29]. This is further compounded by a lack of robustness in attributing audio outputs from third-party voice applications to a specific application and developer in voice assistant ecosystems [25]. Research [1, 30, 35] has also shown that users of voice assistants are not always aware that they are interacting with a third-party voice application (and hence a third-party beyond the voice assistant provider). Therefore, a user may believe that misleading health information, for example, originated from a core provider such as Amazon rather than a third-party voice application, which may enhance the perceived credibility of the information.

3 METHODOLOGY

To explore the issue of misinformation in third-party voice applications, we conducted an empirical study aiming to identify instances of misinformation in live Google Actions and Alexa Skills that are currently accessible to users. Our study methodology comprised four stages Firstly, we defined misinformation in the context of third-party voice applications. Secondly, we gathered data about Actions and Skills to be tested in the study by scraping web data from Amazon and Google marketplaces. Thirdly, we designed a framework and implemented a processing pipeline for interacting with Actions and Skills automatically, including generating the transcript of the interactions. We next discuss these stages in detail.

3.1 Definition of Misinformation

Misinformation has been understood in academic research to include both false information that is distributed unintentionally as a consequence of error, as well as false information that is distributed intentionally for political or other reasons, termed disinformation [8]. Furthermore, misinformation may consist of, though is not synonymous with, 'fake news', that is fabricated or distorted accounts of current events [26, 40]. Lastly, misinformation is often understood to refer not only to inaccurate information but also to biased or subjective information presented as objective truth [32]. For our case study, we define misinformation as potentially harmful information that was either a) verifiably untrue or b) a biased / opinion-based statement presented as fact.

3.2 Action and Skill Sampling

We focus on Actions and Skills referring to 'facts' as a subset of third-party voice applications that might likely contain misinformation and were also relatively easy to test using automated natural language interaction. These types of applications have a simple one-turn or repetitive dialogue structure for interacting with users.

To test Google Actions, data from 590 Actions was scraped from the English language Google Actions marketplace using Selenium Python framework¹. The data includes the name, the URL ID, the average user rating, and the description of the Action. For Alexa Skills testing, we used the data from the UK Alexa Skills marketplace made available by Edu et al. [17] (38,707 Skills). We extracted from the scraped data 5,796 Skills containing the word 'facts' in their title. The dataset includes sample invocation utterances, Skill names, descriptions, and marketplace URL links.

3.3 Natural Language Interaction

3.3.1 Infrastructure. Our testing infrastructure is discussed below.

Natural language interaction agent. We develop a chatbot for interacting with third-party voice applications using Natural Language Processing (NLP). The chatbot is designed to recognise five generic types of output from voice applications, namely:

- requests for a selection ('would you like to go to the wood or to the sea?);
- (2) requests for an instruction ('to begin your workout, say start workout);
- (3) yes/no questions ('do you want to continue?');
- (4) requests for personal information (relating specifically to six types of personal data - name, date of birth/age, gender, address/location, phone number, email); and
- (5) 'open' questions ('what country would you most like to visit?').

The first four of these five generic types of output from voice applications are loosely based on the types of Alexa Skills output identified by [21] (yes/no, instruction, selection and Wh questions). Additionally, we identify the 'open' type of output, asking an apparently open-ended question. Our natural language interaction agent was designed to handle all kinds of output from third-party voice applications.

We use 'bag-of-words' and syntactic features extracted from training data to identify the intent of Skill or Action output. The confidence threshold for intent classification is set to 0 to force the chatbot to identify output from a voice application as one of the five generic input types and return a response. This approach may not generate coherent human-like dialogue in all instances. However, as we aim to collect dialogue data rather than to create human-like interactions for a 'Turing test', we consider that some loss of naturalness might be necessary to maximise the chatbot's ability to respond to a voice application in a multi-turn dialogue rather than exiting the interaction due to low confidence in intent classification. Our chatbot uses dependency parsing to generate responses to Skill or Action outputs identified as instruction or selection requests.

For our implementation, we used an open-source platform called RASA (Bocklisch et al. [7]) designed to support developing generalpurpose chatbots. The intent classification was handled by RASA's built-in transformer-based DIET classifier (Bunk et al. [9]). The Stanza Python toolkit for NLP (Qi et al. [33]) was imported to the

¹https://pypi.org/project/selenium/

Misinformation in Third-party Voice Applications

RASA custom action for dependency parsing of Skill outputs (de Marneffe et al. [15]).

Voice Assistant Clients. The Google Assistant SDK² and the Alexa Voice Service (AVS) Device SDK³ were installed on an Ubuntu virtual machine (version 20.4). Google Assistant SDK uses the gRPC remote procedure call system to communicate with the Google Assistant cloud, and Alexa Voice Service devices communicate with the Alexa cloud using the HTTP/2 multiplexing protocol to exchange JSON-formatted objects between the client and server. We wrote a bespoke Python script to facilitate interaction between the RASA bot and voice assistant clients. Our script reads a list of Skills and Actions to be tested and outputs a file containing a transcription of the dialogue interactions between the bot and the specified applications.

3.3.2 Interaction and Transcript Generation. We used our Python scripts to trigger automated interaction between our RASA bot and target Actions and Skills and to generate transcripts of these dialogue interactions. For Google Actions, the transcripts are based on text-based interaction with Google Assistant SDK, and for Alexa Skills, the transcripts are based on captions of speech output provided by the AVS Device SDK.

Google Actions. We tested (attempted to interact with) all 590 Actions scraped from the marketplace that referred to 'facts' on their Actions marketplace webpage. We note the possibility that some interactions were with Actions of the same or similar name to the Action that we intended to test.

Alexa Skills. We tested (attempted to interact with) 1,022 Skills sampled randomly from the subset of 5,796 Skills containing the word 'facts' in their title. Not all Skills tested were activated, with some being apparently inactive and others being types of Skills that cannot be enabled via the conversational interface.

3.4 Analysis of Dialogue Transcripts

In theory, our methodology supports using a method for automated recognition of misinformation from dialogue transcripts generated by Actions and Skills. However, we found that state-of-the-art automated fact-checking tools do not transfer well to this domain. For instance, the service Google Fact Check⁴ could not identify any of the instances of misinformation we detected in our manual review detailed next, which suggests a potential lack of effectiveness of such existing, general tools for third-party voice applications. Instances of misinformation were instead identified by manual review of transcripts. Where required, we used independent resources for fact-checking (as detailed in Section 4 below).

4 RESULTS

Table 1 summarises the results obtained after analysing 1,612 voice applications with our framework. Overall we identified 17 instances of harmful content as defined above, of which nine were found in Google Actions and eight in Alexa Skills. Table 2 and Table 3 show details of the Actions and Skills we have identified containing misinformation. We next discuss in detail our results per platform.

Table 1: Case Study Findings Summary

Type of Application	Number Tested	Misinformation Instances
Google Actions	590	9
Alexa Skills	1,022	8
TOTAL:	1,612	17

4.1 Google Actions

Our findings on misinformation in Google Actions are listed in Table 2. Nine instances of misinformation were found from the 590 Google Actions tested, Five of these were untrue and potentially harmful, though probably unintentional, namely one instance of inaccurate legal information ('Facts Corner' Action) and four Actions stating that there are 100 calories per Indian meal. There was one instance of a possibly intentional untrue statement about education in Pakistan ('amazing facts about education' Action). This may represent intentional disinformation linked to a known propaganda campaign.⁵ Two instances of misinformation were subjective statements claiming to be factual, in the 'Christianity Facts' and 'Vietnam Facts' Actions. We further identified an Action that claimed to distinguish genuine from fake news ('TattleGEN' Action).

4.2 Alexa Skills

Our findings on misinformation in Alexa Skills are listed in Table 3. From 1,022 Skills, we found eight Skills containing instances of misinformation. Of these eight Skills, only one, the 'Vegan Facts' Skill, was the Skill that was intended to be activated in testing (i.e. on checking the developer account used for testing after interaction with a Skill, the Skill that had been enabled during testing

⁵https://www.bbc.co.uk/news/world-asia-india-55232432

Name	Dialogue transcript	
Amazing Facts	Action trigger: Talk to amazing facts about education	
About Education	Response: OK. Let's get amazing facts about education.	
	Hi I can tell you amazing facts about education. Wanna know?	
	RASA: ok	
	Response: Children living in Pakistan have no legal right to a free	
	education, and compulsory education only runs between the ages of	
	five and nine. Wanna know more?	
	RASA: ok	
	Response: Indian government provides free education to all children	
	between the age of 6 to 14 years.	
Facts Corner	Response: It's against the law to die in the town of Longyearbyen,	
	Norway.	
Calorie Tips	Response: The Average Indian meals Contains 100 Calories.	
TechMastering	Response: The average Indian meals contains 100 callaries.	
questions		
Calculate Calo-	Response: The Average Indian meals contain 100 Calories.	
rie		
Find Calories	Response: The Average Indian Meal contains 100 calories	
Christianity	Response: Here is your fact about Christianity. Christianity is based	
Facts	on fear.	
Vietnam Facts	Response: OK. Here's Vietnam Facts. HO CHI MINH WAS AND STILL	
	IS A NATIONAL HERO	
Tattle GEN	Response: Hi, welcome to TattleGEN! I can help you stay updated	
	about genuine and fake news.	

Table 2: Google Actions Results (excerpts)

 $^{^{2}} https://developers.google.com/assistant/sdk/guides/service/python/assistant/sdk/guides/service/$

³https://developer.amazon.com/en-US/docs/alexa/avs-device-sdk/overview.html ⁴https://toolbox.google.com/factcheck

CUI '23, July 19-21, 2023, Eindhoven, Netherlands

Mary Bispham, Suliman Kalim Sattar, Clara Zard, Xavier Ferrer-Aran, Jide Edu, Guillermo Suarez-Tangil, and Jose Such

Name	Dialogue transcript	
My Trivia	initalmessage: Open my trivia	
	Alexa: myTrivia. I will ask you 5 questions, try to get as many right as you	
	can. Just say the number of the answer.	
	Question 1. Most interesting job? Tech. 2. Defense. 3. Medicine. 4. Legal.	
	RASA: three	
	Alexa: That answer is wrong. The correct answer is 1: Tech.	
Love Facts	Response: Some people can't feel love. Some individuals who claim never	
	to have felt romantic love suffer from hypopituitarism, a rare disease that	
	doesn't allow a person to feel love.	
Freaky Facts	Response: Two diseases that have been eradicated, Chickenpox and Rinder-	
	pest.	
Vegan Facts	Response: A vegetarian diet can help protect your health and even reverse	
	some diseases, including heart disease.	
Funny Facts	Response: Two diseases have successfully been eradicated, Chickenpox	
	and Rinderpest.	
Teething Facts	Response: The safest approach to teething is to return whatever "remedy"	
	you just purchased, thus avoiding anything that may potentially be harmful.	
Surprising	Response: Teenage suicide is the second cause of death in the state of	
Facts	Wisconsin. do you want to listen another fact.	
Global Warm-	Response: The global warming will completely alter the ocean's conveyer	
ing Facts	belt, which will cause a mini ice age in the Europe.	

Table 3: Alexa Skills Results (excerpts)

had the same marketplace URL ID to the Skill that had been targeted). The other seven Skills were activated in place of the intended Skill by audio request. The Skills identified included one Skill ('Love Facts') providing misleading medical information using non-medical terminology, two Skills ('Freaky Facts' and 'Funny Facts') stating incorrectly that Chickenpox disease has been eradicated, one Skill ('My Trivia') stating subjective opinion as fact, two Skills ('Vegan Facts' and 'Teething Facts') provided health information purporting to be factual that is inconsistent with advice from official sources,⁶ one Skill ('Surprising Facts') providing misleading information on suicide statistics as referring to the population as a whole rather than only one age group, and finally a Skill ('Global Warming Facts') containing unsourced and inaccurate information on climate change (the Skill stated as a definite forecast as consequence that is only stated as a possible consequence of global warming in official sources.⁷). This is concerning in the broader context of reported increases in climate change misinformation.

5 DISCUSSION

5.1 Key Findings

Our findings demonstrate the potential for spreading misinformation via third-party voice applications in voice-assisted ecosystems. As well as being significant in themselves, our findings also indicate the likelihood of a more substantial problem that may increase in the future. It is difficult to determine whether the instances of misinformation we have identified in published Actions and Skills imply a failure of initial security vetting by Amazon and Google or modification of the back-end of applications after publication in the marketplaces. Either way, our findings confirm that the current processes for ensuring the security of third-party voice applications in voice assistant ecosystems are inadequate.

Our work implies a need to develop new security measures for third-party voice applications. Exhaustive testing of third-party

⁷See https://oceanservice.noaa.gov/education/tutorial_currents/05conveyor3.html

voice applications before publication may be inherently infeasible due to the lack of scalability of manual testing and the limitations of automated testing with NLP technologies. Even if applications can be tested exhaustively prior to publication, preventing malicious developers from retrospectively altering the back-end code of their applications after vetting by providers is unlikely to be possible in current voice assistant architectures in which back-end code is hosted on third-party servers. Therefore it may be necessary to develop security measures that monitor the live behaviour of third-party voice applications after they are made available in a voice assistant ecosystem.

5.2 Related Work

Misinformation in Voice Assistant Applications. There has been some work on the actual voice assistants spreading misinformation. This work found instances of health-related misinformation (e.g. regarding vaccinations) [3, 20] in Google Assistant, Alexa, Siri, and Cortana. Note, however, that this research relates to information returned from the core voice assistant rather than third-party voice applications (Skills, Actions) created by external developers.

There has been limited research on *the potential* for third-party voice applications to spread misinformation and other harmful content. Sharevski et al. [37] conducted a user study demonstrating the influence of a pseudo-malicious Alexa Skill ('Malexa') that rewords RSS news feeds before delivering them to a listener. In another paper reporting results of another user study focussing specifically on COVID-19 vaccine information, Sharevski et al. [38] demonstrate the potential for users to be influenced by rewording of some types of information relating to the vaccine by a malicious Alexa Skill. Whilst these studies use simulated experimental set-ups to assess the potential effect of misinformation in third-party voice applications, our work seeks to identify actual instances of such content 'in the wild'.

Misinformation in other domains. There has been a significant effort towards identifying misinformation on various platforms like Twitter [5], YouTube [22], and Facebook.⁸, with topics ranging from health to climate science [11, 28]. Social media providers have often relied on users to report harmful content. However, they have also recognised a need for more proactive measures that enable harmful content to be blocked or quarantined before being made available to users [42]. Various approaches detect misinformation, including expert fact-checking, crowd-sourcing, NLP and machine learning-based methods, and others [4, 13, 23, 24, 31]. However, current approaches do not completely solve the problem of misinformation on social media and other platforms in the wild [6, 44], and we showed before how state-of-the-art approaches like Google Fact Check did not detect any of the misinformation cases we found in this domain.

5.3 Limitations and Future Work

Our case study covered only a small subsection of applications in the Actions and Skills marketplaces. It was also subject to our NLP tools limitations. For example, the RASA bot sometimes made NLP errors in responding to output from voice applications, resulting

⁶Official advice does not make unqualified recommendations to follow a vegetarian diet, or advise against the use of teething remedies as such — see https://www.bhf.org.uk/informationsupport/heart-matters-magazine/nutrition/ ask-the-expert/plant-based-diets and https://www.nhs.uk/conditions/baby/babys-development/teething/tips-for-helping-your-teething-baby/ respectively

⁸https://about.fb.com/news/2020/02/online-content-regulation/

Misinformation in Third-party Voice Applications

in premature termination of the interaction. Furthermore, our detection of misinformation was based on transcriptions of audio outputs from third-party voice applications rather than the audio outputs, which may be incomplete. Given these limitations, it is assumed that the amount of potentially harmful information outputted by third-party voice applications is greater in reality than what is shown in our current findings. Therefore, we could consider our findings as a *lower bound* of the actual problem and the tip of the iceberg.

Future work should focus on the automated detection of misinformation and other harmful outputs in live interactions with third-party voice applications, perhaps using an independent monitoring agent. Furthermore, future work should aim to develop methods for detecting harmful content directly from the audio outputs heard by users rather than relying on transcripts of audio output. As automated processing of speech is potentially faster than human processing ⁹, it may be possible to develop methods for detecting and blocking harmful audio outputs in real-time conversational interactions with third-party voice applications, without affecting the human perception of the timing of the interactions that do not contain misinformation or other harmful outputs. However, such detection and blocking must be transparent and ethically conducted to avoid censorship, as an improperly implemented or misused filter could negatively impact users.

ACKNOWLEDGMENTS

This research was funded by EPSRC under grant EP/T026723/1 and by COMET project TED2021-132900A-I00 from MCIN/AEI /10.13039/501100011033 and the European Union-NextGenerationEU/-PRTR. Guillermo Suarez-Tangil was partially funded by the "Ramon y Cajal" Fellowship RYC-2020-029401-I under MCIN/AEI/10.13039/-501100011033 and ESF "The European Social Fund invests in your future." Any opinions, findings, conclusions, or recommendations expressed herein are those of the authors.

REFERENCES

- Noura Abdi, Kopo M Ramokapane, and Jose Such. 2019. More than smart speakers: security and privacy perceptions of smart home personal assistants. In *Fifteenth* Symposium on Usable Privacy and Security ({SOUPS} 2019). 451–466.
- [2] Noura Abdi, Xiao Zhan, Kopo M Ramokapane, and Jose Such. 2021. Privacy Norms for Smart Home Personal Assistants. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 1–14.
- [3] Emily Couvillon Alagha and Rachel Renee Helbing. 2019. Evaluating the quality of voice assistants' responses to consumer health questions about vaccines: an exploratory comparison of Alexa, Google Assistant and Siri. BMJ health & care informatics 26, 1 (2019), e100075.
- [4] Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2021. A Survey on Multimodal Disinformation Detection. arXiv preprint arXiv:2103.12541 (2021).
- [5] Sotirios Antoniadis, Jouliana Litou, and Vana Kalogeraki. 2015. A model for identifying misinformation in online social networks. In OTM Confederated International Conferences" On the Move to Meaningful Internet Systems". Springer, 473–482.
- [6] Agathe Balayn, Jie Yang, Zoltan Szlavik, and Alessandro Bozzon. 2021. Automatic Identification of Harmful, Aggressive, Abusive, and Offensive Language on the Web: A Survey of Technical Biases Informed by Psychology Literature. ACM Transactions on Social Computing (TSC) 4, 3 (2021), 1–56.
- [7] Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. 2017. Rasa: Open source language understanding and dialogue management. arXiv preprint arXiv:1712.05181 (2017).

- [8] Étienne Brown. 2021. Regulating the spread of online misinformation. In The Routledge Handbook of Political Epistemology. Routledge, 214–225.
- [9] Tanja Bunk, Daksh Varshneya, Vladimir Vlasov, and Alan Nichol. 2020. Diet: Lightweight language understanding for dialogue systems. arXiv preprint arXiv:2004.09936 (2020).
- [10] Long Cheng, Christin Wilson, Song Liao, Jeffrey Young, Daniel Dong, and Hongxin Hu. 2020. Dangerous skills got certified: Measuring the trustworthiness of skill certification in voice personal assistant platforms. In Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security. 1699–1716.
- [11] Travis Coan, Constantine Boussalis, John Cook, and Mirjam Nanko. 2021. Computer-assisted detection and classification of misinformation about climate change. (2021).
- [12] Miles C Coleman. 2021. Leveraging the Rhetorical Energies of Machines: COVID-19, Misinformation, and Persuasive Labor. *Human-Machine Communication* 3, 1 (2021), 2.
- [13] Botambu Collins, Dinh Tuyen Hoang, Ngoc Thanh Nguyen, and Dosam Hwang. 2021. Trends in combating fake news on social media-a survey. *Journal of Information and Telecommunication* 5, 2 (2021), 247–266.
- [14] Henry Kudzanai Dambanemuya and Nicholas Diakopoulos. 2021. Auditing the Information Quality of News-Related Queries on the Alexa Voice Assistant. Proceedings of the ACM on Human-Computer Interaction 5, CSCW1 (2021), 1–21.
- [15] Marie-Catherine de Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics* 47, 2 (2021), 255–308.
- [16] Jide Edu. 2022. Assessing and Measuring the Privacy Practices of Voice Assistant Applications. Ph. D. Dissertation. King's College London.
- [17] Jide Edu, Xavier Ferrer-Aran, Jose Such, and Guillermo Suarez-Tangil. 2021. SkillVet: Automated Traceability Analysis of Amazon Alexa Skills. *IEEE Transactions* on Dependable and Secure Computing (2021).
- [18] Jide Edu, Xavier Ferrer-Aran, Jose Such, and Guillermo Suarez-Tangil. 2022. Measuring Alexa Skill Privacy Practices across Three Years. In Proceedings of the ACM Web Conference 2022 (Virtual Event, Lyon, France) (WWW '22). Association for Computing Machinery, New York, NY, USA, 670–680. https://doi.org/10. 1145/3485447.3512289
- [19] Jide Edu, Jose Such, and Guillermo Suarez-Tangil. 2020. Smart home personal assistants: a security and privacy review. ACM Computing Surveys (CSUR) 53, 6 (2020), 1–36.
- [20] John Ferrand, Ryli Hockensmith, Rebecca Fagen Houghton, and Eric R Walsh-Buhi. 2020. Evaluating Smart Assistant Responses for Accuracy and Misinformation Regarding Human Papillomavirus Vaccination: Content Analysis Study. *Journal of Medical Internet Research* 22, 8 (2020), e19018.
- [21] Zhixiu Guo, Zijin Lin, Pan Li, and Kai Chen. 2020. Skillexplorer: Understanding the behavior of skills in large scale. In 29th {USENIX} Security Symposium ({USENIX} Security 20). 2649–2666.
- [22] Eslam Hussein, Prerna Juneja, and Tanushree Mitra. 2020. Measuring misinformation in video search platforms: An audit study on YouTube. Proceedings of the ACM on Human-Computer Interaction 4, CSCW1 (2020), 1–27.
- [23] Bahruz Jabiyev, Jeremiah Onaolapo, VT Burlington, Gianluca Stringhini, and Engin Kirda. [n. d.]. Game of FAME: Automatic Detection of FAke MEmes. ([n. d.]).
- [24] Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. 2021. Toward Automated Factchecking: Developing an Annotation Schema and Benchmark for Consistent Automated Claim Detection. *Digital Threats: Research and Practice* 2, 2 (2021), 1–16.
- [25] Deepak Kumar, Riccardo Paccagnella, Paul Murley, Eric Hennenfent, Joshua Mason, Adam Bates, and Michael Bailey. 2018. Skill squatting attacks on Amazon Alexa. In 27th {USENIX} Security Symposium ({USENIX} Security 18). 33–47.
- [26] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.
- [27] Christopher Lentzsch, Sheel Jayesh Shah, Benjamin Andow, Martin Degeling, Anupam Das, and William Enck. 2021. Hey Alexa, is this Skill Safe?: Taking a Closer Look at the Alexa Skill Ecosystem. In 28th Annual Network and Distributed System Security Symposium (NDSS 2021). The Internet Society.
- [28] Stephan Lewandowsky. 2021. Climate Change Disinformation and How to Combat It. Annual Review of Public Health 42 (2021), 1-21.
- [29] Fei Li and Hosagrahar V Jagadish. 2014. Constructing an interactive natural language interface for relational databases. *Proceedings of the VLDB Endowment* 8, 1 (2014), 73–84.
- [30] David Major, Danny Yuxing Huang, Marshini Chetty, and Nick Feamster. 2021. Alexa, Who Am I Speaking To?: Understanding Users' Ability to Identify Third-Party Apps on Amazon Alexa. ACM Transactions on Internet Technology (TOIT) 22, 1 (2021), 1–22.
- [31] Alexandre Maros, Jussara M Almeida, and Marisa Vasconcelos. 2021. A Study of Misinformation in Audio Messages Shared in WhatsApp Groups. In Multidisciplinary International Symposium on Disinformation in Open Online Media.

⁹https://www.forbes.com/sites/anniebrown/2021/08/08/voice-ai-technology-ismore-advanced-than-you-might-think/

CUI '23, July 19-21, 2023, Eindhoven, Netherlands

Mary Bispham, Suliman Kalim Sattar, Clara Zard, Xavier Ferrer-Aran, Jide Edu, Guillermo Suarez-Tangil, and Jose Such

Springer, 85-100.

- [32] Maria D Molina, S Shyam Sundar, Thai Le, and Dongwon Lee. 2021. "Fake news" is not simply false information: a concept explication and taxonomy of online content. *American behavioral scientist* 65, 2 (2021), 180–212.
- [33] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. arXiv preprint arXiv:2003.07082 (2020).
- [34] William Seymour, Mark Coté, and Jose Such. 2023. Legal Obligation and Ethical Best Practice: Towards Meaningful Verbal Consent for Voice Assistants. In 2023 CHI Conference on Human Factors in Computing Systems. ACM, 285.
- [35] William Seymour and Jose Such. 2023. Ignorance is Bliss? The Effect of Explanations on Perceptions of Voice Assistants. Proceedings of the ACM on Human-Computer Interaction 7, CSCW1 (2023), 1–24.
- [36] William Seymour and Max Van Kleek. 2021. Exploring Interactions Between Trust, Anthropomorphism, and Relationship Development in Voice Assistants. Proceedings of the ACM on Human-Computer Interaction 5, CSCW2 (2021), 1–16.
- [37] Filipo Sharevski, Peter Jachim, Paige Treebridge, Audrey Li, Adam Babin, and Christopher Adadevoh. 2021. Meet Malexa, Alexa's malicious twin: Malwareinduced misperception through intelligent voice assistants. *International Journal* of Human-Computer Studies 149 (2021), 102604.
- [38] Filipo Sharevski, Anna Slowinski, Peter Jachim, and Emma Pieroni. 2021. "Hey Alexa, What do You Know About the COVID-19 Vaccine?"–(Mis) perceptions of

Mass Immunization Among Voice Assistant Users. arXiv preprint arXiv:2105.07854 (2021).

- [39] Dan Su, Jiqiang Liu, Sencun Zhu, Xiaoyang Wang, and Wei Wang. 2020. " Are you home alone?"" Yes" Disclosing Security and Privacy Vulnerabilities in Alexa Skills. arXiv e-prints (2020), arXiv-2010.
- [40] Edson C Tandoc Jr, Zheng Wei Lim, and Richard Ling. 2018. Defining "fake news" A typology of scholarly definitions. Digital journalism 6, 2 (2018), 137–153.
- [41] Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, et al. 2021. Sok: Hate, harassment, and the changing landscape of online abuse. (2021).
- [42] Stefanie Ullmann and Marcus Tomalin. 2020. Quarantining online hate speech: technical and ethical perspectives. *Ethics and Information Technology* 22, 1 (2020), 69–80.
- [43] Dawei Wang, Kai Chen, and Wei Wang. 2021. Demystifying the Vetting Process of Voice-controlled Skills on Markets. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 5, 3 (2021), 1–28.
- [44] Liang Wu, Fred Morstatter, Kathleen M Carley, and Huan Liu. 2019. Misinformation in social media: definition, manipulation, and detection. ACM SIGKDD Explorations Newsletter 21, 2 (2019), 80–90.