Leakage of Sensitive Information to Third-Party Voice Applications

MARY BISPHAM, King's College London, UK
CLARA ZARD, King's College London, UK
SULIMAN SATTAR, King's College London, UK
XAVIER FERRER-ARAN, King's College London, UK
GUILLERMO SUAREZ-TANGIL, IMDEA Networks Institute, Spain
JOSE SUCH, King's College London, UK

In this paper we investigate the issue of sensitive information leakage to third-party voice applications in voice assistant ecosystems. We focus specifically on leakage of sensitive information via the conversational interface. We use a bespoke testing infrastructure to investigate leakage of sensitive information via the conversational interface of Google Actions and Alexa Skills. Our work augments prior work in this area to consider not only specific categories of personal data, but also other types of potentially sensitive information that may be disclosed in voice-based interactions with third-party voice applications. Our findings indicate that current privacy and security measures for third-party voice applications are not sufficient to prevent leakage of all types of sensitive information via the conversational interface. We make key recommendations for the redesign of voice assistant architectures to better prevent leakage of sensitive information via the conversational interface of third-party voice applications in the future.

Additional Key Words and Phrases: voice assistants, conversational interface, voice applications, privacy and security

ACM Reference Format:
Mary Bispham, Clara Zard, Suliman Sattar, Xavier Ferrer-Aran, Guillermo Suarez-Tangil, and Jose Such. 2022. Leakage of Sensitive Information to Third-Party Voice Applications. In . ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

This paper focuses on threats to privacy arising from the disclosure of sensitive information via the conversational interface of third-party voice applications. Providers of voice assistants such as Google Assistant and Amazon Alexa have facilitated the development of increasing numbers of voice applications created by third-party developers that are available on their platforms. Growth in the number of English-language Google Actions was estimated at 340% in 2019[1], whereas the number of available Alexa Skills was recently given as over 100,000.[2] There has been increasing concern of threats to users' privacy from third-party voice applications [1, 7–9, 11, 15]. The porous nature of the conversational interface poses a potential threat to privacy in the context of third-party voice applications in particular. There have been comparable concerns of threats to privacy in relation to the providers of core voice assistants such as Google Assistant and Amazon Alexa. However, in the context of third-party applications, privacy concerns are amplified. As noted by Edu et al. [7], developers of third-party applications do not usually receive a financial reward for their work,

[1]https://voicebot.ai/2020/01/19/google-assistant-actions-grew-quickly-in-several-languages-in-2019-match-alexa-growth-in-english/
[2]https://www.tomsguide.com/uk/round-up/best-alexa-skills

and thus have reduced accountability to users by comparison to commercial providers of core voice assistants. The problem of threat to privacy from third-party voice applications is further compounded by a lack of robustness in attribution of audio outputs from third-party voice applications. Research [2, 12] has shown that users of voice assistants are not always aware that they are interacting with a third-party developer, as opposed to the core voice assistant provider. In fact, when the information flows through the voice assistant ecosystem as made apparent to them, users tend to exhibit very different privacy preferences over third-party voice applications [3].

Whilst there has been some prior work on requests for personal information via the conversational interface of third-party voice applications (e.g., see Guo et al. [11]), this has been limited to identifying requests for specific categories of personal data, and has not considered vulnerability to leakage of sensitive information more broadly. In this paper, we seek to fill the gaps in prior work on leakage of sensitive information via the conversational interface of third-party voice applications. We conduct an empirical study on leakage of sensitive information via the conversational interface of live third-party voice applications in the Google Assistant and Amazon Alexa marketplaces (termed Google Actions and Alexa Skills respectively), using a bespoke testing infrastructure. We identify new types of sensitive information leakage to third-party voice applications not covered in prior work. We provide an assessment of the implications of the findings from our empirical study with regard to the effectiveness of currently available privacy and security measures for third-party voice applications, and provide recommendations for a possible new approach in preventing sensitive information leakage to third-party voice applications.

2 THE CHALLENGE OF PREVENTING INFORMATION LEAKAGE VIA THE CONVERSATIONAL INTERFACE

Preventing leakage of sensitive information via the conversational interface of third-party voice applications remains an outstanding challenge that has yet to be effectively addressed. A key reason for this is that third-party voice applications may run in a remote server not necessarily controlled by the voice assistant provider. Unlike smartphone apps, third-party voice applications are not installed locally, but are instead hosted on servers controlled by the third-party developer. This prevents malicious developers from installing malicious code on user devices, but limits the scope of security testing of third-party voice applications, and also hampers detection of modifications to the back-end code of third-party voice applications. Whereas the availability of source code for mobile apps facilitates security testing of apps using both static and dynamic analysis, security testing of third-party voice applications is limited to dynamic black-box testing (see Wang et al. [16]). In third-party voice applications, the attack surface shifts from the user device to the conversational interface. Whilst third-party voice applications do usually need to pass an individual vetting process before being made publicly available, such vetting processes have been shown to be inadequate; for example, authors in [6] showed that the Amazon Skill certification process can be subverted by crafting mock policy-violating Skills.

3 SENSITIVE INFORMATION LEAKAGE IN GOOGLE ACTIONS AND ALEXA SKILLS

To investigate concrete instances of sensitive information leakage in live third-party voice applications, we implement a testing infrastructure that uses automated natural language interaction to mimic users' voice interactions with third-party voice applications. Our testing infrastructure consists of client agents for commercial voice assistants Google Assistant and Amazon Alexa installed on a virtual machine that link to the Google Assistant and Amazon Alexa cloud environments (the Google Assistant SDK[3] and the Alexa Voice Service (AVS) Device SDK[4]), and a chatbot that interacts

[3]https://developers.google.com/assistant/sdk/guides/service/python
[4]https://developer.amazon.com/en-US/docs/alexa/avs-device-sdk/overview.html

Table 1. Empirical Study Findings Summary

3rd-party application	No. tested	Personal data requests (no. auto-detected)	Other sensitive requests (no. auto-detected)
Google Actions	1,642	20 (13)	5 (2)
Alexa Skills	1,007	4 (2)	3 (0)
TOTAL:	2,649	24 (15)	8 (2)

with Google Actions and Alexa Skills via the client agents. We use an open-source platform called RASA (Bocklisch et al. [5]) designed to give support in developing general-purpose chatbots. Our chatbot uses natural language processing (NLP) to recognise five generic types of output from voice applications, namely requests for selection (*'would you like to go to the woods or to the sea?*), requests for instruction (*'to begin your workout, say start workout'*), yes/no questions (*'do you want to continue?'*), requests for personal information (relating specifically to six types of personal data: 1) *name*, 2) *date of birth/age*, 3) *gender*, 4) *address/location*, 5) *phone number*, and 6) *email*), and '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 Guo et al. [11]. We train our chatbot to identify the five generic types of output from Actions and Skills with a machine-learning classifier available on the RASA platform using a set of training data. Our testing infrastructure is designed with the primary objective of sustaining dialogue interactions with third-party voice applications to the fullest extent possible, and generating transcripts of these interactions that can then be analysed for various purposes.

We use our testing infrastructure to interact with a subset of Google Actions and Alexa Skills, and review the outputs of our testing process in order to identify instances of sensitive information leakage to Actions and Skills. We define sensitive information as consisting either of a request for one of the six specific categories of personal data recognised by our chatbot, or else a request for other types of information that would not typically be in the public domain. We first check the dialogue transcripts generated in the testing process for instances of interactions that the chatbot has detected as a request for personal information. We discard any false positives of requests for personal information detected by the chatbot, i.e. interactions that the bot responds to as a request for personal information, but that does not contain a request for any of the six categories of personal data or for any other information fitting our definition of sensitive information. We then manually review the remaining transcripts, in order to detect any requests for sensitive information that have not been detected automatically by the chatbot.

The results from using our testing infrastructure to interact with a subset of Google Actions and Alexa Skills are summarised in Table 1. These include findings on requests by Actions and Skills for the specific categories of personal data considered in the study, as well as findings on requests for other types of sensitive information. The number of requests for each type of information that were detected automatically by our chatbot is also given. For requests for personal data in the six specific categories, our chatbot was able to automatically identify 15 out of a total of 24 requests across both Actions and Skills (63%). For requests for other types of sensitive information, our chatbot was able to automatically identify two out of a total of eight requests across both Actions and Skills (25%).

Four examples of our findings of sensitive information leakage in Google Actions are shown in Table 2. One of these is a request for a specific type of personal data (name), whereas the other three are new types of sensitive information that were not anticipated in the design of our testing infrastructure.

Table 2. Examples of Sensitive Information Leakage

3rd-party application (type, marketplace URL ID)	Data Type	Dialogue transcript
divorce lawyer (Google Action, 00000082ad61202f	Personal Data (name)	Action trigger: Talk to divorce lawyer Google Assistant: OK. Getting divorce lawyer. Hi My name is Jannet, divorce lawyer. To serve you better, can I have your name please? RASA: That is personal information that I don't want to disclose
McDonalds Apply Thru (Google Action, 00000015922949b0)	Other Sensitive Data (employment)	Action trigger: Talk to McDonald's Apply Thru Google Assistant: Alright. Getting McDonald's Apply Thru. Personal data you provide is being processed by Google and transferred to McDonald's. What country do you want to work in? RASA: I select transferred to McDonald 's Google Assistant: I'm sorry, I didn't understand. Can you please share which country you would like to work in? RASA: Yes
Analyze My Feelings (Google Action, 00000046012f2b87)	Other Sensitive Data (mental health)	Action trigger: Talk to Analyze My Feelings Google Assistant: Alright. Let's get Analyze My Feelings. Greetings! Please explain your situation in your own words RASA: That is personal information that I don't want to disclose
Waifu AI (Google Action, 000000bb805e778c)	Other Sensitive Data (mental health)	Action trigger: Talk to Waifu AI Google Assistant: OK. Let's get Waifu AI. Good day! What can I do for you today? The mic will be left open for you to continue to respond. We suggest replying to what she says and she will continue the conversation. RASA: I select she will continue the conversation Google Assistant: How do you feel when you say that? RASA: Ok, please do Google Assistant: Very interesting. [...] RASA: Ok, ok Google Assistant: Let's change focus a bit... Tell me about your family. RASA: Ok, focus a bit

4 A NEED FOR REDESIGN OF VOICE ASSISTANT ARCHITECTURES?

Our findings on live Actions and Skills in the Alexa and Google Assistant marketplaces provide further confirmation of the vulnerability of users to compromise of sensitive information via the conversational interface of third-party voice applications, and of the inadequacy of current privacy and security measures to prevent this. Furthermore, our findings indicate that the issue of sensitive information leakage may be broader than has been assumed in prior work to date, extending beyond specific categories of personal data to other types of sensitive interactions that may be even more difficult to prevent. As discussed above, our testing infrastructure was able to detect requests for specific categories of personal data automatically in most cases, but was not capable of detecting other types of sensitive information leakage in the majority of instances. Instead, these other types of sensitive information were identified by manual review of dialogue transcripts. In the real-world context, relying on manual review for vetting of third-party voice applications is impractical for reasons of scale, implying a need to improve capabilities for automated vetting. Our work suggests that automated vetting might be improved to some extent by training classifiers to detect not only requests for specific categories of personal data, but also indicators of other types of sensitive information exchange between users and application (such as information on employment as in the 'McDonald's Apply Thru' Action).

However, whilst providing some directions for improvement to vetting processes, our findings also suggest that efforts to improve current vetting processes may face some inherent limitations with respect to their ability to prevent all types of sensitive information leakage. A primary reason for this is that in some instances, the sensitivity of interactions with a third-party voice application might depend not on the outputs from the application, but on the responses of users to the output. For example, if in response to a question such as that from the Waifu AI Action of 'how do you feel when you say that?', a user provides a generic answer such as 'ok' or 'fine', their interaction with the Action might not be considered sensitive. However, if the user replies with a detailed account of their emotional state, it might be considered sensitive. Similarly, different users are likely to provide very different responses to the open-ended request by the 'Analyze my Feelings' Action to 'explain your situation in your own words'. Vetting processes for prevention of sensitive information leakage therefore need not only to detect potentially problematic outputs from third-party voice applications, but also to anticipate potentially compromising user responses (this is in contrast to the development of vetting processes to prevent harmful outputs from voice applications that do not relate to sensitive information leakage, such as output of misinformation or hate speech, that require consideration of outputs only). User responses to outputs from third-party voice applications are likely to be challenging if not impossible to predict for all individual users. This implies a limitation not only for the development of automated vetting processes, but also even for manual vetting, setting aside the fact that reliance on manual vetting is unlikely to be feasible at scale. Even human testers of voice applications might not be able to assess what information other users might provide in response to open-ended questions.

It follows that some types of sensitive information leakage may remain persistently difficult to detect within current voice assistant architectures, without applying a filter that is so broad that it unduly restricts legitimate interactions with third-party voice applications. Therefore, a more fundamental restructuring of architectures for third-party voice applications may be required. One possibility for this is suggested by the concept of digital ephemerality. This concept has also been developed in the context of 'ephemeral' human-to-human messaging apps such as Snapchat, which claim to store no record of communications via the app [10, 13]. A comparable concept of 'transience' has also been proposed for mobile apps [4]. Digital ephemerality might seem especially appropriate as a privacy measure for the context of third-party voice applications, in mimicking the ephemeral nature of real-world spoken communications between humans [14]. A possibility for consideration might be the development of a new 'Snapchat-type' architecture for voice assistants that implements digital ephemerality for interactions between users and third-party voice applications via the conversational interface. In such an architecture, the 'perimeter' of control of third-party voice applications might be redrawn to include back-end code/servers, with the back-end code of third-party voice applications hosted on servers controlled by a voice assistant provider or trusted third party security provider. This would enable the voice assistant or third-party security provider to delete dialogue transcripts after every interaction between a user and a third-party voice application, with the developer of the application never having been given access to these. The application of digital ephemerality in the context of human-machine interaction would clearly have some differences to its application in platforms designed for human-human interaction, such as Snapchat. In the latter case, both of the (human) interaction partners store messages in their memory and can extract value (enjoyment or information) from them after they have been deleted from the digital platform. In human-machine interaction, on the other hand, the implementation of digital ephemerality would mean that developers of third-party voice applications would lose the ability to extract value from messages after they have been consumed. The possible effects of this on the motivation of some third-party developers to create voice applications would need to be considered. However, developers of third-party voice applications would still be able to access any user information legitimately required for the functionality of their applications via a manual permissions API that is separate from the conversational interface.

5 CONCLUSION

Our work has confirmed the potential for compromise of personal data via the conversational interface of third-party voice applications, and also identified new types of sensitive information that might be leaked to third-party developers via the conversational interface that have not been considered in prior work. Based on our findings, we conclude that it may ultimately not be possible to fully prevent leakage of sensitive information via the conversational interface of third-party voice applications in current voice assistant architectures. We recommend that future work should consider possibilities for redesign of voice assistant architectures to facilitate implementation of the concept of digital ephemerality for voice-based interactions with third-party applications.

Acknowledgments

This research was funded by EPSRC under grant EP/T026723/1 and the "Ramon y Cajal" Fellowship RYC-2020-029401-I.

REFERENCES

[1] 2022. SkillDetective: Automated Policy-Violation Detection of Voice Assistant Applications in the Wild. In *31st USENIX Security Symposium (USENIX Security 22)*. USENIX Association, Boston, MA. https://www.usenix.org/conference/usenixsecurity22/presentation/young

[2] Noura Abdi, Kopo Ramokapane, and Jose Such. 2019. More than Smart Speakers: Security and Privacy Perceptions of Smart Home Personal Assistants. In *Fifteenth USENIX Symposium on Usable Privacy and Security (SOUPS 2019)*. 451–466.

[3] Noura Abdi, Xiao Zhan, Kopo Ramokapane, and Jose Such. 2021. Privacy Norms for Smart Home Personal Assistants. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. 558:1–558:14.

[4] Raquel Alvarez, Jake Levenson, Ryan Sheatsley, and Patrick McDaniel. 2019. Application Transiency: Towards a Fair Trade of Personal Information for Application Services. In *International Conference on Security and Privacy in Communication Systems*. Springer, 47–66.

[5] Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. 2017. Rasa: Open source language understanding and dialogue management. *arXiv preprint arXiv:1712.05181* (2017).

[6] 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.

[7] 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).

[8] Jide Edu, Xavier Ferrer-Aran, Jose Such, and Guillermo Suarez-Tangil. 2022. Measuring Alexa Skill Privacy Practices across Three Years. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*. ACM, 670–680.

[9] Jide Edu, Jose Such, and Guillermo Suarez-Tangil. 2021. Smart home personal assistants: a security and privacy review. *ACM Computing Surveys (CSUR)* 53, 6 (2021), 1–36.

[10] Ryan G Ganzenmuller. 2014. Snap and destroy: Preservation issues for ephemeral communications. *Buff. L. Rev.* 62 (2014), 1239.

[11] 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.

[12] 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.

[13] Esther Shein. 2013. Ephemeral data. *Commun. ACM* 56, 9 (2013), 20–22.

[14] Oren Soffer. 2016. The oral paradigm and Snapchat. *Social Media+ Society* 2, 3 (2016), 2056305116666306.

[15] Jose Such. 2017. Privacy and Autonomous Systems. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. 4761–4767.

[16] 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.