Undergraduate Thesis Prospectus

Voice Privacy: Preventing Alexa from Eavesdropping

(technical research project in Computer Science)

The Struggle for Control of Personal Data

(sociotechnical research project)

by

Matthew Hancock

November 2, 2020

technical project collaborators:

Gabriel Simmons
Tu Le
Danny Huang

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.
Matthew G. Hancock

Technical advisor: Yuan Tian, Department of Computer Science

STS advisor: Peter Norton, Department of Engineering and Society

General Research Problem

How can unwanted interactions with personal voice assistants be prevented?

1 in 4 Americans own a smart home device such as a Google Home, Amazon Alexa, or Apple Homepod with a personal voice assistant (Bitterly 2019). Factor in smartphones with built-in voice assistants, and a lot of people are near devices that are listening to them at all times. Although they stay in standby until hearing the trigger phrases “Hey Siri,” “Okay Google,” or “Alexa,” accidental activations are frequent. They can occur at the worst times: during medical consultations or confidential interviews, for example. Users would not have to entrust large tech companies with their personal data if they could prevent the companies from accessing it.

Voice Privacy: Preventing Alexa from Eavesdropping

How can trends and patterns in Amazon Alexa voice history aid owners in tracking recorded conversations and deleting unwanted voice data?

I am doing undergraduate research for my capstone project this year. It is through the Computer Science department at UVA, my technical capstone advisor is professor Yuan Tian. I am collaborating with Gabriel Simmons, another Undergraduate researcher. Tu Le, a PHD candidate at UVA, is our PHD mentor who is guiding us along as he works on his own research. Danny Huang, an Assistant Professor at NYU, is providing additional guidance to us and Tu.

My technical project involves giving users a tool to download the usage data of Amazon Alexa devices and build classifiers to help users manage voice history. The tool allows users to download conversations heard by Alexa to their local machine, track who

uses Alexa for what purposes, search through specific interactions, and auto-delete all conversation history or just sensitive history. Once our tool is out, we will analyze what data participants are deleting and why to better help decide what data should be auto-deleted. We hope to get this tool released by the end of the spring semester.

Our goal is to put privacy into the forefront of users' minds when they use voice assistants. A constraint is that our extension only works with Amazon Alexa, we can't help manage voice data from Google Assistant, Siri, and other voice assistants. The state of the art is going to the voice history webpage on Amazon, looking at and listening to all interactions with Alexa, keeping wanted voice interactions on Amazon's servers, and individually deleting unwanted interactions (Amazon). There are currently two webpages that need to be visited if the user wants all data amazon is keeping on them. Our method of creating the tool to manage voice history is creating a google chrome extension. We are coding it in JavaScript. This introduces another constraint: users will have to use google chrome to participate. Right now, we are only using the chrome extension on our local machines, it hasn't been released to the public yet. We test and experiment using the extension on our own personal Alexa voice history. To get information on what participants are deleting, we will prompt them to explain why they are deleting voice interactions when done through the extension.

We hope to complete the chrome extension by the end of the spring. Once finished, the extension will be released to the public, and our team will collect usage data for a few months. Once we have collected enough data, we hope to write a conference paper on our findings.

The Struggle for Control of Personal Data

In the U.S. since 2010, how have privacy advocates attempted to limit data collection by voice assistants?

Interactions with voice assistants provide a great resource to the technology companies that make them. With voice histories, tech companies can better understand what users say, target marketing to them, and sell data to advertisers. When asked what they listen to and store in servers, voice assistants tell users they awake only to a trigger phrase ("Alexa," Hey Siri," or "Okay Google"); they then offer a link to the tech company's privacy webpage. However, voice assistants frequently misinterpret other sounds as the trigger phrase, when they may turn active and record a sensitive conversation. Privacy advocates, lawmakers, and general consumers alike want to know why unwanted conversations are recorded and stored on tech companies' servers, to what extent more conversations are recorded without users' consent, andwhat companies are doing to limit unauthorized data collection and storage. They demand useful devices that also protect users' privacy.

According to Acquisti, Taylor, and Wagman (2016) privacy is an economic value entailing tradeoffs with other values, such as utility. A more secure product may be less usable. For example, a more secure device may require hands-on operation, may require a manually activated microphone, and may not recognize the user's voice. Martin (2016) proposes that online privacy is a voluntary contract between users and data collectors. Users do not give their data away; they trade it in a mutually beneficial exchange. The user of a voice assistant must trust that the tech company records only by user consent, and that user-device interactions are secure. Ferraris et al. (2020), propose a user trust

model in which users trust scores are derived from household size, number of visitors, and types of conversations near the voice assistant and similar variables.The scores indicate which voice assistants suit the user's trust standards.

Participants include the manufacturers of voice assistants: Google, Amazon, and Apple. Google's profit model depends on data collection for ad targeting and ad revenue. Advertisers pay Google to personalize ads to each user. Google Assistant answers many user questions through Google searches. Google claims that Google Assistant keeps no record of voice interactions unless the user prompts it to do so, and that the company sells no data (Google). Amazon profits through purchases on its online marketplace. With its voice assistant, Alexa, users can quickly purchase products from Amazon. Alexa stores all voice interactions on its servers unless the user deletes them (Amazon). Apple profits from hardware sales. Its products' voice assistant, Siri, is built in, or it can be connected to one device from another device. Apple claims it stores no voice history except by user prompt, and that data are stored only to improve product utility (Apple, n.d.).

Privacy advocates are demanding that voice assistants be subject to stricter data collection, storage, and use standards. Echo Kids Privacy (EKP) is a coalition of advocacies, including Center for Digital Democracy, Campaign for a Commercial-Free Childhood, and Parents Across America. Contending that by unintentionally recording children's conversations, Alexa violates the Children's Online Privacy Protection Act (COPPA), EKP submitted a formal complaint to the Federal Trade Commission (EKP, 2019). Some lawmakers seek to regulate the collection of voice data. They favor stricter voice data use standards, such as preventing the sale of voice data for targeted

advertising. (Silvestro et al., 2016). U.S. Senator Chris Coons of Delaware, a Democrat, has objected to Amazon's data collection practices (Huseman 2019).

Malicious third parties can gather voice data surreptitiously. They may sell data to advertisers, they may hold sensitive conversations for ransom, or they may serve foreign intelligence agencies. They can attack voice assistants through wiretapping, through hacking attacks, or by activating assistants without the user's knowledge (Chung et al., 2017).

References

Acquisti, A., Taylor, C., & Wagman, L. (2016). The Economics of Privacy. *Journal of Economic Literature, 54(2),* 442-492. Jstor

Amazon. Alexa and Echo devices are designed to protect your privacy. Amazon.com. https://www.amazon.com/b/?node=19149155011

Apple. Privacy. Apple.com. https://www.apple.com/privacy/.

Bitterly, Kourtney (2019, August 22). 1 In 4 Americans Own a Smart Speaker. What Does That Mean for News? NYT Open. https://open.nytimes.com/how-might-the-new-york-times-sound-on-smart-speakers-3b59a6a78ae3.

Chung, H., Iorga, M., Voas, J., & Lee, S. (2017). Alexa, Can I Trust You? *Computer, 50(9),* 100-104.

EKP (2019, May 9). Echo Kids Privacy. Echo Dot Kids Edition Violates COPPA. https://www.echokidsprivacy.com

Ferraris, D., Bastos, D., Fernandez-Gago, C., & El-Moussa, F. (2020). A trust model for popular smart home devices. *International Journal of Information Security*. https://doi.org/10.1007/s10207-020-00519-2

Google. Assistant Privacy & Security Features - Google Safety Center. safety.google. https://safety.google/intl/en_us/assistant/?utm_source=google.

Martin, K. (2015). Understanding Privacy Online: Development of a Social Contract Approach to Privacy. *Journal of Business Ethics*, 137(3), 551-569

Silvestro, M., & Black, J. (2016). "Who Am I Talking To?"— The Regulation of Voice Data Collected by Connected Consumer Products. *Business Law Today, 1*, 1-4.

Huseman, B. (2019, June 28). Response to Concerns over Amazon's privacy and Data Collection Practices [Letter to Christopher A. Coons]. United States Senate, Washington, District of Colombia.