Undergraduate Thesis Prospectus

Parallelization of Speech Recognition Model Evaluation

(technical research project in Computer Science)

Perception of Privacy in Virtual Assistants: Consumers vs. Corporations

(sociotechnical research project)

by

Teagan Le

October 27, 2022

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.

Teagan Le

STS advisor: Peter Norton, Department of Engineering and Society

General research problem

How is consumer artificial intelligence being developed and marketed?

"Artificial intelligence" is marketed to consumers for many applications, including self-driving cars (Waymo, n.d.), speech recognition, virtual assistants, home surveillance, and consumer robotics (Lau et al., 2018, p. 1; Tritschler, 2021). This set of technologies is becoming an inescapable part of technological society: it is used by Google for YouTube's recommendation system (Goodrow, 2021), which is used by over one billion people every month (YouTube, 2013), and over 100 million Alexa devices have been sold (Bohn, 2019). The limits of the field continue to be pushed by researchers and companies hoping to apply the technology to their products. These companies engage in marketing of their AI products to consumers, in order to get people to buy them. The intersection of these efforts presents an interesting engineering and sociotechnical area of research.

Parallelization of Speech Recognition Model Evaluation

How can distributed computing be used to decrease execution time for machine learning workflows?

In the Computer Science department, I understand that the capstone is to write a technical report for CS 4991 detailing my work during my internships. I worked on distributed computing systems for automated speech recognition for a popular virtual assistant. The goal of the project was to reduce the runtime of a workflow that evaluated models used for speech recognition for metrics such as word and sentence error rate. To this end, I parallelized a section of the workflow by splitting the work over multiple computers, while ensuring the behavior was preserved.

Before I implemented my solution, the workflow was partially parallelized: it ran several instances of the model and ran evaluations in parallel, but it collected the results from each instance to perform further analysis, which presented a bottleneck in the system. My work involved extending the parallelization to some of these analyses, so that the results are collected after all of the work is done, and no work is done after the parallel results are aggregated. This was accomplished using the organization's internally-developed framework for distributed machine learning workflows.

Perception of Privacy in Virtual Assistants: Consumers vs. Corporations

In North America, how are consumers, tech companies, and privacy advocates competing to determine the limits of permissible data collection by virtual assistants?

Virtual assistants are developed by several vendors, including Google, Apple, and Amazon. Amazon SVP Dave Limp revealed that "more than 100 million devices with Alexa on board have been sold" (Bohn, 2019). With so many devices sold, there are persistent questions about how user data is utilized by virtual assistant providers (Augustin et al., 2022, p. 310; Perez, 2019). The data that is collected by these virtual assistants can be extensive; an article from Reuters on Alexa's data collection states:

> Such information can reveal a person's height, weight and health; their ethnicity (via clues contained in voice data) and political leanings; their reading and buying habits; their whereabouts on any given day, and sometimes whom they have met. (Kirkham & Dastin, 2021)

This information can be used for many purposes, like improving the machine learning models that interpret human speech into text (Amazon, 2022; Depuy et al., 2022), or to provide users

with personalized advertising (Google, n.d. b; Tuohy, 2022). In one case, the recorded data from virtual assistants was used as evidence in a criminal investigation (Heater, 2017).

Virtual assistants have had a controversial track record with regards to their data collection and usage. Specifically, Amazon Alexa has had many headlines criticizing its data practices (Lynskey, 2019; Perez, 2019). On the other hand, Google Assistant appears to have had fewer public incidents regarding its data practices. Amazon and Google, along with other vendors of virtual assistant services, constitute the first major participant group in this research question. Amazon markets its Alexa virtual assistant as wanting "to make your life easier, more meaningful, and more fun by helping you voice control your world" (Amazon, n.d. b). These two companies represent much of the market share for virtual assistants (Bishop, 2021; Bohn, 2019), so they are the most relevant participants. All of these companies must reduce people's hesitation due to privacy issues, both real and perceived, to drive sales.

Amazon uses tactics such as an information page about Alexa privacy, claiming that "Amazon designs Alexa and Echo devices with multiple layers of privacy and security" (Amazon, n.d. a). Google has several similar pages, where they address topics such as (Google, n.d. a; n.d. b). Both Amazon and Google have published videos where they explain privacy features on their virtual assistants and smart speakers: Amazon touts its physical security features, such as an electronic microphone disconnect and a camera blind, and both companies emphasize how their smart speakers do not send data unless the "wakeword" is detected on the device (Amazon Alexa, 2020a; 2020b; Google, 2020). Amazon teams have published articles about their work in advancing secure machine learning technology that obfuscates the user data that is used in model training (Depuy et al., 2022). It seems that these companies attempt to allay fears about the security of their products through educational marketing.

The second group of participants are digital privacy activists and researchers who are trying to expose security issues and unethical privacy practices. These include academic researchers, who have found vulnerabilities with the "skills" system of Amazon Alexa (Das & Shipman, 2021), or commercial security researchers, such as those at the security firm Checkmarx (Newman, 2018). This group also includes many advocacy groups that focus on privacy and civil liberties.

One major advocacy group in this domain is the Electronic Frontier Foundation. The EFF claims to be "the leading nonprofit organization defending civil liberties in the digital world" (Electronic Frontier Foundation, n.d.). They say they work for "user privacy, free expression, and innovation through impact litigation, policy analysis, grassroots activism, and technology development," and that their mission is to "ensure that technology supports freedom, justice, and innovation for all people of the world" (Electronic Frontier Foundation, n.d.). The EFF has voiced support for Amazon's decisions surrounding Alexa and various warrants that were placed for data collected as part of Alexa's functions (Williams, 2017). The EFF has also repeatedly criticized and called on Amazon to stop partnering with law enforcement agencies with its Ring cloud-based home surveillance systems (Guariglia, 2019).

Another important advocacy group is Privacy International. They say that they "protect democracy, defend people's dignity, and demand accountability from the powerful institutions who breach public trust" (Privacy International, n.d. a). The actions they take include performing investigations on how data is being collected and used, running awareness campaigns, and proposing "legal and technological frameworks to protect against data exploitation" (Privacy International, n.d. b).

There are other consumer advocacy groups that publish reports on how secure various consumer gadgets are, based on their assessments. One of these groups is the Mozilla Foundation, which advocates for a "movement to realize the full potential of the internet" (Mozilla Foundation, n.d. a). They state their three approaches to this goal as "rallying citizens," "connecting leaders," and "shaping the agenda" (Mozilla Foundation, n.d. a). They perform investigations into possible threats for what they see as a "healthy internet", and publish their research as open-source materials, which is open to anyone who wants to see and use it (Mozilla Foundation, n.d. a). Mozilla says they create their "Privacy Not Included" buyer's guide to help consumers get clear information about the security and privacy practices of various consumer devices (Mozilla Foundation, n.d. b). They claim their methodology is entirely from reading the published information about a product by its vendor, such as the privacy policy and advertising (Mozilla Foundation, n.d. b). Mozilla has published reports on several virtual assistant devices, from Apple, Google, and Amazon. For example, in their review of Amazon's Echo Show smart display, they mention Amazon's policies on data deletion rights, data access granted to third-party developers, and Amazon Alexa's poor privacy track record (Mozilla Foundation, 2021).

The last major group of participants are the users of these virtual assistants. Users' agendas are to spend their money where they believe the privacy tradeoff is fair, and to opine on the privacy issues of these services. These users include journalists, such as Dorian Lynskey, who writes: "people who consider them sinister and invasive (myself included) regard enthusiasts as complacent, while those who find them useful and benign see the sceptics as paranoid technophobes" (2019).

Insights into how consumers perceive the privacy of virtual assistants are provided by market surveys and studies by academic researchers. A study by Microsoft claims that 41% of

users are concerned about privacy (Perez, 2019). Lau, et al. attempted to uncover what factors lead to the adoption or non-adoption of smart speakers, and specifically how perceptions of privacy tied into this (2018, p. 2). They found that non-users "did not trust speaker companies," (Lau et al., 2018, p. 1) and that users "justify their lack of privacy concern based on an incomplete understanding of the privacy risks" (p. 21), and that "both users and non-users exhibit resignation to privacy loss" (p. 22). Augustin, et al. provide an analysis of past studies on the subject, and posit ideas such as "privacy fatigue" to explain user behaviors, as well as recommendations for future studies (2022, p. 1). Research into user forums and social media has found that a large number of people are unaware of any privacy concerns with virtual assistants, and even when users know about privacy concerns they do not necessarily adapt their behaviors (Augustin et al., 2022, p. 318). Liao, et al. studied the privacy policies used in third-party applications for Alexa and Google Assistant, and found that many apps did not have a valid privacy policy (2020, p. 1). Liao, et al. also found that most users ignore privacy policies, and many users are unaware of what data is being collected (2020, p. 865).

References

Amazon. (n.d. a). *Alexa Privacy – Learn how Alexa works*. Amazon.com. https://www.amazon.com/Alexa-Privacy-Hub/b?ie=UTF8&node=19149155011

Amazon. (n.d. b). *Amazon Alexa – Learn what Alexa can do*. Amazon.com. https://www.amazon.com/b?node=21576558011&ref=pe_alxhub_aucc_en_us_IC_NV_HUB_HP

Amazon. (2022, January 28). How Amazon protects customer privacy while making Alexa better. *About Amazon.* https://www.aboutamazon.com/news/devices/how-amazon-protects-customer-privacy-while-making-alexa-better

Amazon Alexa. (2020a, January 31). *Learn about Alexa privacy features and controls* [Video]. YouTube. https://www.youtube.com/watch?v=l38YWC5vDaI

Amazon Alexa. (2020b, April 13). *Alexa Privacy: How Does Alexa Work?* [Video]. YouTube. https://www.youtube.com/watch?v=esMOQgDMAeo

Augustin, Y., Carolus, A., & Wienrich, C. (2022, June 16). Privacy of AI-Based Voice Assistants: Understanding the Users' Perspective. *Lecture Notes in Computer Science, 13337*, 309-321. Springer. https://doi.org/10.1007/978-3-031-05014-5_26

Bishop, T. (2021, August 4). Amazon maintains big lead over Google and Apple in U.S. smart speaker market, new study says. *GeekWire*. https://www.geekwire.com/2021/amazon-maintains-big-lead-google-apple-u-s-smart-speaker-market-new-study-says/

Bohn, D. (2019, January 4). Exclusive: Amazon says 100 million Alexa devices have been sold. *The Verge*. https://www.theverge.com/2019/1/4/18168565/amazon-alexa-devices-how-many-sold-number-100-million-dave-limp

Das, A., & Shipman, M. (2021, March 4). Study Reveals Extent of Privacy Vulnerabilities With Amazon's Alexa. *NC State News*. https://news.ncsu.edu/2021/03/alexa-skill-vulnerabilities/

Depuy, C., Dhamala, J., & Gupta, R. (2022, April 28). Advances in trustworthy machine learning at Alexa AI. *Amazon Science*. https://www.amazon.science/blog/advances-in-trustworthy-machine-learning-at-alexa-ai

Electronic Frontier Foundation. (n.d.). *About EFF*. Electronic Frontier Foundation. https://www.eff.org/about

Goodrow, C. (2021, September 15). On YouTube's recommendation system. *YouTube Blog.* https://blog.youtube/inside-youtube/on-youtubes-recommendation-system/

Google. (n.d. a). *How Google Assistant works with your data - Google Assistant Help*. Google Support. https://support.google.com/assistant/answer/11091015

Google. (n.d. b). *Protecting Your Google Assistant Privacy*. Google Safety Center. https://safety.google/assistant/

Google. (2020, January 7). *Privacy On Google Assistant* [Video]. YouTube. https://www.youtube.com/watch?v=ZaqZcDOoi-8

Guariglia, M. (2019, August 8). Amazon's Ring Is a Perfect Storm of Privacy Threats. *Electronic Frontier Foundation.* https://www.eff.org/deeplinks/2019/08/amazons-ring-perfect-storm-privacy-threats

Heater, B. (2017, March 7). Amazon hands over Echo data in Arkansas murder trial. *TechCrunch*. https://techcrunch.com/2017/03/07/amazon-echo-murder/

Kirkham, C., & Dastin, J. (2021, November 19). A look at the intimate details Amazon knows about us. *Reuters*. https://www.reuters.com/technology/look-intimate-details-amazon-knows-about-us-2021-11-19/

Lau, J., Zimmerman, B., & Schaub, F. (2018, November 1). Alexa, Are You Listening?: Privacy Perceptions, Concerns and Privacy-seeking Behaviors with Smart Speakers. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1-31. https://doi.org/10.1145/3274371

Liao, S., Wilson, C., Cheng, L., Hu, H., & Deng, H. (2020, December). Measuring the Effectiveness of Privacy Policies for Voice Assistant Applications. *Annual Computer Security Applications Conference (ACSAC 2020)*, 856-869. https://doi.org/10.1145/3427228.3427250

Lynskey, D. (2019, October 9). 'Alexa, are you invading my privacy?' – the dark side of our voice assistants. *The Guardian.* https://www.theguardian.com/technology/2019/oct/09/alexa-are-you-invading-my-privacy-the-dark-side-of-our-voice-assistants

Mozilla Foundation. (n.d. a). *What we do*. Mozilla Foundation. https://foundation.mozilla.org/en/what-we-do/

Mozilla Foundation. (n.d. b). *Why we made *privacy not included*. Mozilla Foundation. https://foundation.mozilla.org/en/privacynotincluded/about/why/

Mozilla Foundation. (2021, November 8). Amazon Echo Show | Privacy & security guide. **Privacy Not Included*. https://foundation.mozilla.org/en/privacynotincluded/amazon-echo-show/

Newman, L. H. (2018, April 25). Turning an Echo Into a Spy Device Only Took Some Clever Coding. *WIRED*. https://www.wired.com/story/amazon-echo-alexa-skill-spying/

Perez, S. (2019, April 24). 41% of voice assistant users have concerns about trust and privacy, report finds. *TechCrunch*. https://techcrunch.com/2019/04/24/41-of-voice-assistant-users-have-concerns-about-trust-and-privacy-report-finds/

Privacy International. (n.d. a). *Challenging Corporate Data Exploitation*. Privacy International. https://privacyinternational.org/strategic-areas/challenging-corporate-data-exploitation

Privacy International. (n.d. b). *The future we want*. Privacy International. https://privacyinternational.org/demand/the-future-we-want

Tritschler, C. (2021, September 28). Meet Astro, a home robot unlike any other. *About Amazon.* https://www.aboutamazon.com/news/devices/meet-astro-a-home-robot-unlike-any-other

Tuohy, J. P. (2022, April 28). Report shows that Amazon uses data from Alexa smart speakers to serve targeted ads. *The Verge*. https://www.theverge.com/2022/4/28/23047026/amazon-alexa-voice-data-targeted-ads-research-report

Waymo. (n.d.). *Waymo Driver – Waymo*. Waymo. https://waymo.com/waymo-driver/

Williams, J. (2017, March 9). EFF Applauds Amazon For Pushing Back on Request for Echo Data. *Electronic Frontier Foundation.* https://www.eff.org/deeplinks/2017/03/eff-applauds-amazon-pushing-back-request-echo-data

YouTube. (2013, March 21). YouTube Hits a Billion Monthly Users. *YouTube Official Blog.* https://blog.youtube/news-and-events/onebillionstrong/