Replicating Apple's Private Cloud Compute for AI: Privacy and Performance Trade-offs
(Technical Paper)

Trust in the Machine: An Actor-Network Analysis of Privacy in Apple Intelligence
(STS Paper)

A Thesis Prospectus Submitted to the
Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements of the Degree
Bachelor of Science, School of Engineering

Riley Immel
Fall 2024

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

Signature ____ Date __________
Riley Immel

Technical Advisor: Angela Orebaugh, Department of Computer Science

STS Advisor: Richard D. Jacques, Ph.D., Department of Engineering & Society

Introduction

Artificial Intelligence has become one of the fastest-growing technologies of modern time and can provide many benefits such as generating images, answering questions, writing code, and more based on its learning from existing data (Busch, 2023). However, AI has various limitations and concerns centered around it. Some of these concerns are centered around bias as if the training data is itself biased, it can lead to a biased AI even if "…no bigotry was programmed into the system…" (Dr. Varsha, 2023). Another major concern as AI becomes more integrated into society is protection of user data and user privacy. Apple, the creator of the iPhone, has announced that coming Fall 2024, it is releasing what it calls "Apple Intelligence", a privacy-preserving AI model integrated into Apple devices. Notably, Apple Intelligence will be built into, and a main feature, of the latest operating system for iPhones, of which Apple sold 217.7 million in 2018[1].

Beyond this integration, however, a key selling point of Apple Intelligence is that it will draw on the user's personal data to be more in tune with the user. An emerging concern that stems from this however, is the privacy of user data as Apple Intelligence is not always able to fully complete tasks on the user's device and as a fall back, will make use of their new cloud-based server called "Private Cloud Compute", which will require sending highly sensitive user data to the server. To better understand the limitations of on-device vs. cloud processing and the ecosystem Apple is creating, my technical project will aim to replicate this ecosystem with the use of a Raspberry Pi, a desktop computer, and custom AI to perform tasks during which metrics will be recorded and analyzed to understand what the limitations are of on-device processing and how necessary a cloud server is. For my STS paper, I will be researching the current state of the

[1] Apple stopped reporting iPhone sales in Q4 2018, thus all other numbers are simply estimates

art of data privacy and its intersection with AI as well as looking at the relationships between consumer, user privacy, and large corporations like Apple through the Actor-Network Theory. My STS research will closely align with my technical project, as the privacy methods discovered through my research will be implemented in my technical project and their effect on performance evaluated to better understand their trade-offs and real-world impact.

Technical Discussion

As it currently exists, AI is a resource-intensive technology and thus is limited in what devices can run it. To obtain the best and fastest results, advanced and focused hardware is required. Of course, not all use cases of AI require such hardware or need to be the absolute best and fastest and thus in some cases AI is able to function on more general-purpose hardware. This dynamic is the backbone of the new ecosystem that Apple is implementing with the launch of their Apple Intelligence feature. As a result of the current state of hardware and AI, Apple is limiting what devices will be able to use the feature, for two primary reasons. The first being that not all of their devices have the necessary hardware to use it as they were not designed with such a feature in mind. As a result, only the two most recent iPhone models, 15 Pro and 16/16 Pro, will have access as they have the hardware to support it. The second reason, which is more relevant to my project and STS research, is that while the ecosystem has the PCC to manage more intensive processing, in the world of data privacy, cloud computing is considered less secure than on-device processing. On-device processing is when the necessary processing to manage a request can be entirely done on the requesting device. This ability to manage the request all alone is why on-device processing is considered to be the best in terms of data privacy; the user data is never exposed to the outside world like it is in off-device processing during transmission to the cloud server. In the case of Apple Intelligence, this means that no sensitive user data ever leaves the device in the handling of the request; unless of course, the PCC is needed to manage the request.

An obvious drawback of on-device processing, however, is that the device itself needs to have the resources necessary to manage all processing, which is the reason behind Apple limiting which devices get the feature. Currently, most AI models use off-site or cloud servers for handling requests. However, this comes with drawbacks that include increased response times due to factors associated with internet traffic, connection speed, and the need to transport data between requester and server, which is inherently less safe as the data is leaving a controlled environment.

In the case of Apple Intelligence, the ecosystem will feature both on and off device processing with it favoring on-device whenever possible and off-device as a fallback when necessary. This setup becomes a concern as the data used in processing can be extremely sensitive user data and thus potentially exposed when off-device processing is used. To alleviate these concerns, Apple makes the claim that their PCC server is revolutionary in terms of privacy and does not store nor make user data accessible to anyone but the user themselves. This type of ecosystem is noteworthy as it is the first of its kind, which raises concerns over Apple's promises of its privacy.

The goal of the technical project is to replicate this new ecosystem Apple is implementing for their Apple Intelligence to better understand multiple aspects of it. To achieve the project, I plan on using a Raspberry Pi as my "iPhone," a desktop computer (which will have a larger pool of resources than the Pi) as my "PCC," and some form of custom AI model to perform tasks that include image identification or voice processing. Once a model is found or constructed that aligns with the needs of the project, I will begin to collect data about different performance metrics. I plan to collect data on multiple fronts including running the same task on vs off device and examining metrics such as total time to receive and manage the request, quality and/or accuracy of the output, and resources used to manage the request. I plan to determine what the upper limits of the Pi's ability to manage requests is and then find what a quality upper bound is when traits like response time and resource usage are desired to be lower. I then plan to collect the aforementioned data across multiple experiments in which different forms of privacy are implemented when using the "server" for processing. These are likely to be methods including no privacy, encryption, and others that I discover during my STS research.

Through this project I hope to better understand what is possible in terms of on vs off device processing while ensuring data privacy, and data privacy's effect on performance metrics. After arriving at my conclusions on these points, I hope to have a better understanding of Apple's new ecosystem in terms of what the ratio of on vs off device processing could look like.

STS Discussion

Artificial intelligence is increasingly interwoven into daily life, offering benefits but also raising concerns about privacy and security. Apple's latest AI development, Apple Intelligence, claims to balance the user experience with robust privacy protections. By favoring on-device processing for tasks when possible and offloading more intensive processing to their PCC when necessary, Apple promises that user data remains secure and private. This claim is particularly significant in light of rising consumer awareness and concern over data privacy, as well as regulatory scrutiny on tech giants that includes Apple, Google, and Amazon.

The primary stakeholders in this research are the consumers, who are the most affected by Apple's handling of personal data. Apple itself is a large stakeholder, both as the developer and data processor, and has a vested interest in maintaining user trust and ensuring compliance with privacy regulations. Regulatory bodies are also critical stakeholders, as they are tasked with protecting consumer rights and enforcing transparency. Lastly, privacy advocates and watchdog organizations monitor companies' adherence to privacy standards, often advocating for more stringent controls on data processing practices. In addition, developers and evaluators of new platforms will also have an increased interest in Apple Intelligence as Apple has included their PCC in their bounty program with up to a $1 million reward for any actor who can misuse it and demonstrate this to Apple.

In terms of physical artifacts, Apple's devices and PCC servers represent key components of the privacy ecosystem. These are the platforms through which user data is collected, processed, and in some cases stored. Non-physical artifacts include the software powering Apple Intelligence and Apple's public privacy statements and promises. Together, these artifacts shape consumer perceptions, trusts, and beliefs, making Apple's privacy claims an influential part of its marketing strategy.

The Actor-Network Theory (ANT) will serve as the theoretical framework for my research, as it allows for the analysis of the interactions between human and non-human actors in Apple's ecosystem. ANT is a theory which focuses on human and non-human actors and the connections between them as well as how new actors form as a result of said connections. By using ANT, the relationships between consumers, devices, cloud servers, and Apple itself as they contribute to, or challenge, Apple's privacy claims will be revealed. While ANT has been criticized for downplaying power differences, in this case, it still offers valuable insights into the networked nature of privacy, data processing, and public opinion. By applying ANT, this research will explore how Apple's technologies, privacy policies, and AI models collectively create or undermine the consumer's trust in Apple's privacy promises.

This research is critically important given that it addresses the growing societal concerns about privacy and the ethical use of AI in consumer devices in a rapidly growing technology focused world. Furthermore, Apple's privacy stance is not only shaping user expectations and beliefs but also influencing industry standards which needs to be examined. Understanding these practices is essential for maintaining consumer trust and developing potential regulatory responses and governmental law. Furthermore, by integrating these insights into my technical project, I expect to provide a more comprehensive understanding of privacy-preserving techniques in the realm of data privacy while bridging theory with practical application in AI development.

Conclusion

The technical project—a simulated Apple Intelligence ecosystem using a Raspberry Pi for on-device processing and a desktop as the Private Cloud Compute (PCC) server—aims to demonstrate the practical limits and privacy implications of on-device vs. off-device processing. This project will reveal situations in which off-device processing is required and assess the impact different data privacy methods have on performance metrics like response time and quality.

The STS research will be based on the Actor-Network Theory, which will analyze interactions among key stakeholders in Apple Intelligence that includes consumers, regulatory bodies, and Apple. The research portion aims to shed light on the dynamics of the relationships among the stakeholders and how consumer trust is impacted by human and non-human actors while potentially highlighting areas where more transparency is needed in addition to determining differing forms of data privacy practices.

The overall goal of the technical project and STS research is to better understand the new emerging technologies including AI and to draw conclusions on the ethical nature of technologies like Apple Intelligence and to be more informed about the relationships that exist between consumers, corporations, and other actors involved in these technologies.

References

Apple Intelligence Preview. (n.d.). Apple. Retrieved October 4, 2024, from https://www.apple.com/apple-intelligence/

Apple Statistics (2024). (n.d.). Business of Apps. Retrieved November 8, 2024, from https://www.businessofapps.com/data/apple-statistics/

Busch, K. E. (2023). *Generative Artificial Intelligence and Data Privacy: A Primer* (Internet materials; CRS Report). Congressional Research Service. https://purl.fdlp.gov/GPO/gpo213803

Gutiérrez, J. L. M. (2023). On actor-network theory and algorithms: ChatGPT and the new power relationships in the age of AI. *AI and Ethics*. https://doi.org/10.1007/s43681-023-00314-4

IEEE Electronic Library (IEL) Conference Proceedings. (2021). *2021 International Conference on Artificial Intelligence for Cyber Security Systems and Privacy (AI-CSP)* (Internet materials). IEEE. http://RE5QY4SB7X.search.serialssolutions.com/?V=1.0&L=RE5QY4SB7X&S=JCs&C=TC_046945499&T=marc

IEEE Staff & IEEE Electronic Library (IEL) Conference Proceedings. (2014). *2014 International Conference on Privacy and Security in Mobile Systems (PRISMS)* (Internet materials). IEEE. http://RE5QY4SB7X.search.serialssolutions.com/?V=1.0&L=RE5QY4SB7X&S=JCs&C=TC0001774056&T=marc

iPhone 16 and AI: Everything you missed from Apple's "Glowtime" event. (2024, September 10). Euronews. https://www.euronews.com/next/2024/09/10/new-iphone16-and-apple-intelligence-what-you-missed-from-the-glowtime-event

Latour, B. (1996). On actor-network theory: A few clarifications. *Soziale Welt*, *47*(4), 369–381.

Nasir, N. (2024). *Untangling the Cloud From Edge Computing for the Internet of Things* [University of Virginia, Computer Science - School of Engineering and Applied Science, PHD (Doctor of Philosophy), 2024]. https://doi.org/10.18130/ya3p-vn55

P.   s.   , Dr. V. (2023). How can we manage biases in artificial intelligence systems – A systematic literature review. *International Journal of Information Management Data Insights*, *3*(1), 100165. https://doi.org/10.1016/j.jjimei.2023.100165

Shandilya, S. K., Chun, S. A., Shandilya, S., & Weippl, E. (2018). *Internet of Things Security: Fundamentals, Techniques and Applications* (Internet materials). River Publishers.

https://proxy1.library.virginia.edu/login?url=https://www.taylorfrancis.com/books/9781003338642

Sharma, N. (Computer scientist) (editor), Srivastava, D. (Computer scientist) (editor), & Sinwar, D. (editor). (2024). *Artificial Intelligence Technology in Healthcare: Security and Privacy Issues* (Internet materials). CRC Press. https://proxy1.library.virginia.edu/login?url=https://www.taylorfrancis.com/books/9781003377818

Vaidya, J., Gabbouj, M., & Li, J. (Eds.). (2024). *Artificial Intelligence Security and Privacy: First International Conference on Artificial Intelligence Security and Privacy, AIS&P 2023, Guangzhou, China, December 3–5, 2023, Proceedings, Part I* (Vol. 14509). Springer Nature. https://doi.org/10.1007/978-981-99-9785-5