Earlence Fernandes – Research Statement

I. Overview

I take a broad view of computer systems security. The classical interpretation focuses on the host operating system and network levels. However, computer systems *themselves* have evolved. Today, they are distributed, embedded and capable of learning. Although these evolved computer systems provide benefits in many areas including energy efficiency, productivity, convenience and safety, unless they have the correct security foundations, they amplify traditional computer security threats, giving those threats wider reach and impact and also introduce new classes of threats. My research goal is to ensure that human society will gain the benefits of these evolved computer systems without the security and privacy issues.

To achieve this vision, I use a systems security approach — (1) apply the security mindset to deeply understand the new ways in which these evolved computer systems are vulnerable; and (2) construct principled systems-level defenses that offer end-to-end guarantees. In general, I take particular inspiration from the secure design principles of computer systems, as articulated by Saltzer and Schroeder [21]. My current research agenda focuses on two themes that I believe are critical enablers of the evolved computer systems discussed above.

(1) Controlled sharing in distributed systems. Due to the commoditization of cloud computing, all of our digital data and physical devices are now available on the Internet sitting behind vendor-specific cloud services and APIs (e.g., emails, personal files, messages, health data, smart home and building devices). The main benefit is that users can share their resources with other parties to achieve useful functionality. One of the most popular systems that use controlled sharing are so-called Trigger-action platforms that evolved to support user-specified connections between data and devices. Common examples include controlling a garage door based on location and saving emails of a certain subject to online storage. Controlled sharing protocols (e.g., OAuth) make all this flexible automation possible, yet, our collective experience has shown that these protocols are insufficient from a security standpoint. Fundamentally, they cannot control *how* sensitive access is used once privilege has been granted. Indeed, many of the problems of user resource misuse today are fundamentally because there are no controls on how third-parties use their privileged access. My research goal is to create techniques for controlled sharing in distributed systems where one can obtain strong guarantees on *how* that access is used by a third party. The core insight is to enforce the classic principle of least-privilege by co-designing sharing protocols and the systems that use the shared resources.

(2) ML-powered physical systems. Machine learning has supercharged many areas of computer systems. Simultaneously, the community has established that machine learning is fundamentally vulnerable to many kinds of attacks. This problem is particularly pernicious when we use ML in safety- and security-critical systems like autonomous vehicles or authentication systems in mixed reality. My key insight is that ultimately, machine learning is a piece in a larger computational pipeline and thus, understanding and building robustness *requires* an end-to-end *systems* perspective. Building on this insight, this theme establishes a foundational knowledge about how ML-powered physical systems are vulnerable and lays the groundwork for creating end-to-end robustness.

Impact. I believe that if there are ways to benefit society *after* publishing papers, then it is important to facilitate such positive impact. Earlier in my career, I collaborated with Samsung to fix security issues in their smart home product line. Based on that experience, ideas from my subsequent work in secure smart home platforms (e.g., FlowFence [15]) was adapted into Samsung products. I am currently exploring how defense techniques from my work on secure Internet-scale automation can be applied to products from IFTTT (one of the largest user-centered automation providers). The Stop sign that we used in our physical adversarial example work is on display at the Science Museum in London (and has been added to their permanent collection of artifacts), and helps raise awareness in the general public about emergent security issues.

I often engage with the press when society at large would benefit from awareness about my results — my work on smart home security and physical adversarial perturbations work received widespread coverage, and raised awareness among key stakeholders. For example, I have advised a member of Congress on Internet of Things security, and several US government agencies including the FTC, the National Academies, and the JASONs on security issues related to the deployment of ML systems in critical infrastructure.

II. Research Themes

I will discuss recent research results that focus on two areas: (1) Secure controlled sharing in user-centered automation platforms and (2) Secure ML-powered physical systems. The interested reader may refer to the bibliography for pointers to my work in smart home security [6]–[8], [14], [15], [18], privacy-respecting data analysis systems [16], [17], and my early work in smartphone security [3], [11]–[13], [19], [20], [23].

A. Secure controlled sharing in user-centered automation platforms

The digital and physical resources of people, such as emails, health data, smart home devices, and smart city devices, are now accessible on the Internet. By bringing all these systems online and making them interoperable, system operators enable new functionality and drive efficiencies. The enabler of such useful interconnections is the Internet-scale automation system, whose hallmark is permitting non-programmers to create automations, thus democratizing the bridge between digital and physical resources. Unfortunately, these automation systems are not secure and do not guarantee user privacy – attackers can steal sensitive user data and manipulate resources, including physical ones, at large scale. My thesis is that core issue is one of insufficient controlled sharing. Extant protocols, like OAuth2, are incapable of addressing the core problem of regulating *how* third-parties use their access to sensitive resources once they are authorized. This research theme establishes techniques for controlled sharing of a user's digital and physical resources while respecting the principle of least-privilege in how that access is used. I have started contributing novel controlled sharing protocols in the context of end-user Internet-scale automation systems. I have found that the structure of these systems that make them expressive and widely-used is also beneficial for securing them.

Shared resource access minimization (USENIX Security 2022) [1]. Consider the automation rule from Fig. 1. Based on the OAuth controlled sharing protocol, the automation platform will obtain OAuth tokens to access user emails and a Slack channel. The security issue of this design is twofold: (1) The OAuth tokens allow access to more data than what is needed to run the rule; (2) An attacker can misuse the tokens because the security system cannot control how an entity uses those credentials. In this example, the automation platform gets access to all emails for a user from the specified sender (due to issue 1 above) and if those tokens are leaked (e.g., through a simple implementation flaw), the attacker can use these tokens however they want (due to issue 2 above).

Stepping back, the ideal solution is to have a controlled sharing system that ensures: (1) Anyone who has the authorization tokens may *only use* them in accordance with a user-specified policy; (2) The tokens are least-privileged with respect to the rule. This is a challenging problem to solve in general. First, what is the appropriate access usage policy? Simply asking users is a non-solution. Second, how can we create least-privilege tokens that are specific to computations that a user wants to run with their data and devices?

My work's core insight is that the well-defined structure and semantics of end-user automations lend themselves to automatic techniques for the two challenges above. Specifically, we show how a controlled sharing system can automatically derive an access usage policy directly from the automations that end-users create. Using the derived policy, it shows how one can utilize lightweight program analysis techniques to enforce that the automation platform can only access the user's resources in accordance with the derived policy.

```
let str = NewEmail.Subject
if (str.indexOf('Confidential') ===
↪  -1) {
  Slack.postToChannel.skip()
} else {
  Slack.postToChannel.setMessage(
  NewEmail.Subject
  + ' just received!')
}
```

Fig. 1: Example automation rule that computes on sensitive data, connecting an email service with Slack. It has access to ALL email data from a sender, but only needs the subject line, only when it contains the characters 'Confidential'. The 'skip' function call means that the entire rule is aborted. In that case, the rule does not need access to *any* sensitive email data.

Going back to the example in Fig. 1, we offer the following guarantee, automatically: The automation cloud server (where the user-created automations execute) will be able to use its OAuth tokens to access an email's subject line *only* when the email subject line contains the specified keyword, despite the fact that the tokens can themselves do a lot more than that. This effectively locks down the *use* of those tokens automatically and dynamically.

Cryptographically-enforced controlled sharing (IEEE S&P 2021) [2]. The above project offers one notion of controlled sharing — a third-party with a security token may gain minimal access to

user data, as specified by a user-created automation rule. Although this is a step in the right direction, ultimately, the automation cloud server does get access to plaintext data, that it could potentially misuse. Is there a stronger form of controlled sharing possible where this type of misuse is not mitigated? My encrypted automation system explores this question by adapting tools for computing on encrypted data. The core insight is that the rules are simplistic and stateless compared to general-purpose software. This allows us to tailor cryptographic techniques like Garbled circuits to gain practicality, functionality and efficiency.

Concretely, we designate the automation cloud server as an evaluator of Garbled Cicuits (GCs), while the endpoint sources and sinks of data are trusted (indeed, if they are not, no guarantees are possible). To achieve this design, my work addresses the following open problems: (1) Although from a theoretical perspective, any computable function can be evaluated using GCs, performance matters. Therefore, what types of operations exist in trigger-compute-action rules and can we efficiently execute them using appropriate circuit representations? What innovative circuits do we need for the TAP domain? (2) The automation cloud is fully malicious implying that we must use malicious-secure protocols; however, they can become expensive and introduce multi-round communications between parties which drastically changes the TAP computing paradigm. Is there a design point where we can mitigate these issues while supporting real-world trigger-action rules?

At a high-level, this project found that indeed, it is possible to run automation rules using garbled circuits in the presence of a malicious function evaluator using *only a semi-honest implementation of GCs*. This is possible because the automation environment affords a trusted circuit generator and a trusted party that operates on the results of GC evaluation.

Current and future directions. Based on my experience with the above two projects, I have established that stronger controlled sharing systems are indeed possible to build with good functionality, performance and usability. Looking ahead, I am curious about a more general-purpose solution to controlled sharing. Indeed, one can view the above two systems as point-solutions for a specific style of sharing in distributed systems. One concrete plan is to leverage the recent technological trend of *tailored* trusted computing primitives, as exemplified by the Keystone-RISC-V [9] and Komodo-ARM [5] projects. My high-level concept is to create security tokens that may only be used in minimalist trusted enclaves that support very specific types of computations. These minimalist enclaves are tailored to the specific type of computation being performed with user data. This is contrary to current trusted computing design approaches where they strive for backwards-compatibility that unfortunately leads to very large trusted computing bases running inside enclaves. In the long term, I believe this research theme will lead to a framework of techniques and tools that will help future distributed system designers to enforce strong controlled sharing guarantees to protect users digital and physical resources.

B. Securing ML-powered Physical Systems

AI has super-charged many areas of computer systems, including those that interact with the physical world. For example, self-driving cars use machine learning for perception and authentication systems in mixed reality devices use it to recognize biometrics. My core insight is that these are essentially evolved computer systems that are capable of learning, and thus, securing them end-to-end demands a systems perspective. This theme's goal is to create foundational knowledge about the end-to-end robustness properties of computer systems when they use ML as part of the pipeline. Only then, will we be able to construct appropriate systems-oriented defenses. My current work focuses on the canonical ML-powered physical system: autonomous vehicles. In the future, I plan to expand the investigation to include systems like biometric authentication in mixed reality.

Realistic threat models (CVPR 2018 & 2021, AAAI 2021). The most realistic way to compromise an ML-powered autonomous vehicle is to manipulate its environment. When it senses this corrupted environment, its behavior will change from ideal and the attacker will achieve their goals. My early work showed the first example of how an attacker can throw a few stickers on a Stop sign to trick traffic sign classifiers [4]. However, this opens up several questions, such as: (1) What is the space of physical adversarial examples that are *effective* and *realistic*? (2) What are the end-to-end effects of attacks on machine learning models in a control pipeline?

For the first question on examining the attack space, my goal is to examine the differences between human and machine sensing and understand how attackers can exploit these differences. By establishing these new attack classes, we can begin to think about defenses. As preliminary work, we have recently demonstrated the first *invisible* physical attack that only manipulates the light shining on victim objects rather than the object's underlying texture [22].

This work exploits the fact that humans cannot perceive flickering light above 80 Hz. Fig. 2 shows an example. I contend that these types of attacks are more realistic and practical than current techniques (e.g., attacker flying a drone in front of a victim car or aiming lasers from the side of the road at moving objects with high precision) because the attacker only has to hack a smart LED bulb (e.g., in a tunnel) to achieve their goals.

For the second question on end-to-end effects, my goal is to determine how ML-based attacks affect the larger systems that use ML as part of their processing pipelines. This will help determine what parts of the pipelines are vulnerable and what kinds of defenses need to be in place to avoid the security issues. As preliminary work, we have recently contributed a sequential attack algorithm that can manipulate the behavior of Forward Collision Warning (FCW) systems — a common driver assistance feature available in most cars today. Using techniques from control theory, we demonstrate how an attacker can only compromise vision-related measurements to completely hijack FCW behavior [10]. For example, an attacker can force FCW to show no collision danger when in fact, there is an imminent collision.

With Attack Signal Without Attack Signal

Fig. 2: Example of an invisible physical attack: Images as seen by a human (without border) and as captured by a camera (in black border) with the attack signal (left two images) and without (right two images). The image without the attack signal is classified as coffee mug, while the image with the attack signal is classified as perfume. The attack is robust to camera orientation, distance, and ambient lighting.

Unfortunately, we discovered that attacking an end-to-end system such as FCW requires us to make threat model assumptions that I would characterize as unrealistic (concurrent work in the community has come to a similar conclusion). Specifically, current attacks with end-to-end effects depend on a number of hard-to-achieve factors to align perfectly. This makes them fragile and thus, limits their impact. Building defenses for such types of threat classes is insufficient because they do not adequately capture real-world attacks. Thus, defining real-world threat models is very much an open problem and my work is establishing the foundations in this space.

Current and future directions. Motivated by the above results that show an existence proof of potential security issues, I am pushing our understanding of practical physical attacks along two directions: (1) Investigate other types of physical attacks that rely on differences between human and machine sensing. For example, global shutter cameras are vulnerable to electromagnetic interference — could an attack direct specially-crafted RF signals and manipulate what the camera is seeing? (2) Apply practical physical attacks to end-to-end self-driving pipelines to understand global robustness properties. The second direction will help focus the defense efforts of the community on realistic threat models. Longer term, I plan on exploring how ML security issues affect biometric authentication that is used in augmented reality systems, as this is another crucial and emerging area where ML is an integral part of the larger computer system. I have recently received funding from Facebook on related issues in authentication for mixed reality.

REFERENCES

[1] Y. Chen, M. Alhanahnah, A. Sabelfeld, R. Chatterjee, and **E. Fernandes**, "Practical Data Access Minimization in Trigger-Action Systems," in *Proceedings of the 31st USENIX Security Symposium*, 2022.

[2] Y. Chen, A. Chowdhury, R. Wang, A. Sabelfeld, R. Chatterjee, and **E. Fernandes**, "Data Privacy in Trigger-Action Systems," in *Proceedings of the 42nd IEEE Symposium on Security and Privacy (S&P)*, 2021.

[3] M. Conti, B. Crispo, **E. Fernandes**, and Y. Zhauniarovich, "CRePE: A System for Enforcing Fine-Grained Context-Related Policies on Android," *IEEE Transactions on Information Forensics and Security (TIFS)*, 2012.

[4] K. Eykholt, I. Evtimov, **E. Fernandes**, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust Physical-World Attacks on Deep Learning Visual Classification," in *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[5] A. Ferraiuolo, A. Baumann, C. Hawblitzel, and B. Parno, "Komodo: Using verification to disentangle secure-enclave hardware from software," in *Proceedings of the 26th Symposium on Operating Systems Principles*, ser. SOSP '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 287–305. [Online]. Available: https://doi.org/10.1145/3132747.3132782

[6] W. He, M. Golla, R. Padhi, J. Ofek, M. Dürmuth, **E. Fernandes**, and B. Ur, "Rethinking access control and authentication for the home internet of things (iot)," in *27th USENIX Security Symposium (USENIX Security 18)*. Baltimore, MD: USENIX Association, 2018, pp. 255–272. [Online]. Available: https://www.usenix.org/conference/usenixsecurity18/presentation/he

[7] W. He, V. Zhao, O. Morkved, S. Siddiqui, **E. Fernandes**, J. Hester, and B. Ur, "Sok: Context sensing for access control in the adversarial home iot," in *2021 IEEE European Symposium on Security and Privacy (EuroS P)*, 2021, pp. 37–53.

[8] Y. Jia, Q. A. Chen, S. Wang, A. Rahmati, **E. Fernandes**, Z. M. Mao, and A. Prakash, "ContexIoT: Towards Providing Contextual Integrity to Appified IoT Platforms," in *21st Network and Distributed Security Symposium*, 2017.

[9] D. Lee, D. Kohlbrenner, S. Shinde, K. Asanović, and D. Song, "Keystone: An open framework for architecting trusted execution environments," in *Proceedings of the Fifteenth European Conference on Computer Systems*, ser. EuroSys '20. New York, NY, USA: Association for Computing Machinery, 2020. [Online]. Available: https://doi.org/10.1145/3342195.3387532

[10] Y. Ma, J. Sharp, R. Wang, **E. Fernandes**, and X. Zhu, "Sequential Attacks on Kalman Filter-based Forward Collision Warning Systems," in *35th AAAI Conference on Artificial Intelligence (AAAI)*, Feb. 2021.

[11] **E. Fernandes**, A. Aluri, A. Crowell, and A. Prakash, "Decomposable Trust for Android Applications," in *2015 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, 2015.

[12] **E. Fernandes**, Q. A. Chen, J. Paupore, G. Essl, J. A. Halderman, Z. M. Mao, and A. Prakash, "Android UI Deception Revisited: Attacks and Defenses," in *Proceedings of the 20th International Conference on Financial Cryptography and Data Security (FC)*, 2016.

[13] **E. Fernandes**, B. Crispo, and M. Conti, "FM 99.9, Radio Virus: Exploiting FM Radio Broadcasts for Malware Deployment," *IEEE Transactions on Information Forensics and Security (TIFS)*, 2013.

[14] **E. Fernandes**, J. Jung, and A. Prakash, "Security Analysis of Emerging Smart Home Applications," in *Proceedings of the 37th IEEE Symposium on Security and Privacy (S&P)*, 2016.

[15] **E. Fernandes**, J. Paupore, A. Rahmati, D. Simionato, M. Conti, and A. Prakash, "FlowFence: Practical Data Protection for Emerging IoT Application Frameworks," in *Proceedings of the 25th USENIX Security Symposium*, 2016.

[16] **E. Fernandes**, O. Riva, and S. Nath, "My OS Ought to Know Me Better: In-app Behavioural Analytics as an OS Service," in *15th Workshop on Hot Topics in Operating Systems (HotOS XV)*, 2015.

[17] **E. Fernandes**, O. Riva, and S. Nath, "Appstract: On-The-Fly App Content Semantics with Better Privacy," in *Proceedings of the 22nd ACM Annual International Conference on Mobile Computing and Networking (MobiCom)*, 2016.

[18] A. Rahmati, **E. Fernandes**, K. Eykholt, and A. Prakash, "Tyche: A risk-based permission model for smart homes," in *Proceedings of the 3rd IEEE CyberSecurity Development Conference (SecDev)*, 2018.

[19] G. Russello, M. Conti, B. Crispo, and **E. Fernandes**, "MOSES: Supporting Operation Modes on Smartphones," in *Proceedings of the 17th ACM Symposium on Access Control Models and Technologies (SACMAT)*, 2012.

[20] G. Russello, B. Crispo, **E. Fernandes**, and Y. Zhauniarovich, "YAASE: Yet Another Android Security Extension," in *3rd IEEE Conference on Privacy, Security, Risk and Trust (PASSAT)*, 2011.

[21] J. Saltzer and M. Schroeder, "The protection of information in computer systems," *Proceedings of the IEEE*, vol. 63, no. 9, pp. 1278–1308, 1975.

[22] A. Sayles, A. Hooda, M. Gupta, R. Chatterjee, and **E. Fernandes**, "Invisible perturbations: Physical adversarial examples exploiting the rolling shutter effect," in *In proceedings of CVPR 2021*.

[23] Y. Zhauniarovich, G. Russello, M. Conti, B. Crispo, and **E. Fernandes**, "MOSES: Supporting and Enforcing Security Profiles on Smartphones," *IEEE Transactions on Dependable and Secure Computing (TDSC)*, 2014.