Research Statement

My research is inherently multidisciplinary: it lies at the intersection of computer security and privacy with machine learning (ML) and human-computer interaction, in the fields of adversarial ML and human factors in security and privacy.

ML algorithms enable transformative technologies (e.g., face recognition, cancer diagnosis, and autonomous driving), and are becoming increasingly ubiquitous in computer systems. Introducing ML algorithms into systems can potentially increase the systems' attack surface. Namely, attackers can attempt to exploit vulnerabilities in ML algorithms to cause harm. Early work in adversarial ML has sought to characterize the vulnerability of ML algorithms to attacks, but the types of harm that attackers can cause and whether attackers can cause harm under real-world constraints remain largely unknown. My work in the area aims to answer these questions by studying practical attacks against ML-based systems at deployment time (including for face recognition and malware detection) and defenses against them.

Another entity in security and privacy, crucial yet often forgotten, is the user. Understanding users (their behaviors, needs, perceptions, preferences, ...) is helpful for developing user-centered defenses that would be widely adopted and used as intended, thus enhancing users' security and privacy. My research in human factors in security and privacy aims to draw data-driven insights to enhance our understanding of users and improve user-centered defenses. Specifically, I use quantitative and qualitative methods from data and social sciences and apply them to datasets of various sizes (collected via carefully designed user studies with tens or hundreds of participants, or by industry collaborators with access to observational data pertaining to tens or hundreds of thousands of users) to learn more about users, and explore how to personalize systems to address users' specific needs.

Adversarial ML

Despite their high performance in various challenging tasks, ML algorithms in general, and deep neural networks (DNNs) in particular, are vulnerable to attacks. Amongst the different types of attacks, evasion attacks at inference time—attackers misleading ML algorithms via strategically manipulated inputs, commonly known as adversarial examples—are potentially the most pernicious.

Early work on adversarial examples focused primarily on *imperceptible* attacks, in settings where the objective is to find perturbations of small magnitudes (specifically, as measured by L_p-norms) and the attackers have complete control over inputs. While keeping the perturbations' magnitudes small aims to ensure the attacks' imperceptibility to humans, my work has shown that this is neither necessary (e.g., imperceptible image rotations require large perturbation magnitudes) nor sufficient (our user-study participants could notice perturbations of small magnitudes) for producing successful attacks [8]. Moreover, and perhaps more importantly, early attacks were impractical. For example, an attacker attempting to fool face recognition in practice may only control her physical appearance, and cannot precisely control the exact value of any pixel in the face image being classified.

To fill the gap, my thesis work studied evasion attacks that pose a practical threat to state-of-the-art ML systems, including ones for face recognition and malware detection [10, 11, 12]. For example, my research pioneered physical-world attacks against ML systems, by demonstrating how to produce

eyeglass frames that, when printed and worn, can be used by attackers to *dodge* face recognition (i.e., get arbitrarily misclassified as other individuals) or *impersonate* specific individuals. These physically realizable eyeglass frames had to satisfy multiple objectives, including attaining smooth transitions between neighboring pixels, colors that can be printed by a commodity printer, and robustness against volatile imaging conditions (e.g., different poses or expressions). I developed a systematic method to produce eyeglasses that satisfy these desired objectives, and demonstrated successful attacks against state-of-the-art DNNs for face recognition. The techniques I proposed inspired impactful research on physical-world attacks against street-sign and object recognition [1, 4].

To help mitigate the risks of adversarial examples, my thesis work also proposed a novel defense, n-ML, inspired by n-version programming [7]. While n-version programming relies on independent programs to detect unexpected inputs that trigger bugs or exploit vulnerabilities, n-ML uses an ensemble of n diversified DNNs to detect adversarial examples. At inference time, n-ML classifies inputs by a vote of the DNNs, where each is trained via *topology manipulation*—a new method I proposed to train DNNs to classify benign inputs correctly and to (mis)classify adversarial examples differently than one another, according to a certain specification. Consequently, the DNNs mostly agree on benign inputs, and disagree on adversarial examples, thus creating an opportunity to detect the latter. When evaluated in various application domains (including face and street-sign recognition), we found that n-ML: *1)* roughly retained the benign accuracies of state-of-the-art DNNs; *2)* outperformed prior defenses at detecting adversarial examples in several settings; and *3)* had lower inference-time overhead than the majority of prior methods for detecting adversarial examples we tested.

Future Work in Adversarial ML

The long-term goal of my research in adversarial ML is to explore practical attacks in new domains, and build on insights learned from attacks to inform efficient and effective defenses.

Unexplored Sensors The vast majority of the work on practical adversarial examples on ML systems have considered systems that use typical sensors, such as RGB cameras and microphones. Systems that rely on other sensors, such as near infrared and depth cameras, are widely used in practice (e.g., in Windows Hello and iPhones), yet their susceptibility to adversarial examples has not been rigorously studied. I would like to study practical attacks against such systems, and explore whether certain sensors, or a combination thereof, can result in better security.

Industrial Control Systems (ICSs) I am interested in exploring the impact of adversarial ML on the security of ICSs, such as water treatment systems and smart grids, which are becoming increasingly network connected and reliant on ML (e.g., for forecasting consumption). Adversarial examples targeting ICSs may have severe consequences (e.g., poisoning of drinking water), but also need to adhere to certain physical constraints (e.g., readings of a water-pressure sensor have to be in certain ranges). I have successfully devised algorithms to produce adversarial examples that adhere to constrains in my prior work, and believe that similar approaches can be useful when studying ICSs' security.

Novel Defenses A primary goal behind studying attacks is informing the development of more secure systems. So far, the main focus of research has been on improving ML algorithms' robustness against imperceptible adversarial examples (i.e., ones produced by perturbations with small L_p-norms), which have limited practical implications. Moreover, prior defenses often result in models that have relatively low inference throughput, and thus are unsuitable for certain critical tasks. Differently from prior work, I would like to: *1)* take advantage of *physical-world simulations* (e.g., 3D rendering of faces) to allow

rapid prototyping and evaluation of defenses against physical-world attacks, as these are often hindered by the sizeable effort and time consumed by manual evaluation; *2)* explore how to take advantage of adversarial examples to defend against adversaries that can be faithfully approximated with ML models (e.g., adversarial examples against an adversary performing user profiling can inform how to protect users' privacy); and *3)* study the *robustness and efficiency tradeoffs* that can be achieved by various models and using different feature types to devise robust models that provide fast inference for specialized tasks (e.g., classifying URLs as malicious or benign in real-time for millions of users, as done by large computer-security vendors). Preliminary work that I have done in each of these directions has shown promising findings.

Human Factors in Security and Privacy

Security and privacy researchers have long aimed to understand users' security and privacy needs, how they use security and privacy tools, and the factors that lead user to behave insecurely or affect their willingness to share their data. Without such understanding, security and privacy tools that may seem effective in theory are unlikely to be adopted by users, or could even be misused by users, thus leading to counterproductive outcomes. To help improve our understanding, I studied factors that impact users' valuations of their personal data [14], users' perceptions of computer-security issues [9] and the risks that end-user programming services of Internet-of-Things devices pose to their data [2], and the effectiveness of educational tools and activities at helping vulnerable populations avoid scams [3].

Despite prior efforts, most security and privacy tools follow one-size-fits-all designs (e.g., ad-blockers and anti-viruses do not adapt to users), and our understanding of users' behaviors and preferences remains mostly limited to users from WEIRD (western, educated, industrialized, rich, and democratic) cultures. However, as my work has shown, users from certain cultures (e.g., Japan) exhibit less secure behavior than others [6]. Moreover, users varying in their ages, levels of expertise, and other characteristics, face different online risks, and require different types of interventions to address their security and privacy needs (e.g., some users are best served by aggressive interventions that prioritize their security, while others do not) [3, 9, 13]. Hence, user-centered tools that are tailored to individuals' needs can help us provide enhanced security and privacy protections to users.

Toward moving away from the one-size-fits-all mindset, I have explored ways to leverage users' behavior, preferences, and other, available, contextual information to build personalized security and privacy tools. In the realm of security, I showed that users' Internet-browsing patterns and responses to survey questions assessing their security behavior can help predict exposure to malicious websites ahead of time [13]. Via a collaboration with KDDI, a large mobile phone provider from Japan, we surveyed more than 20,000 cell phone users, and obtained three months' worth of HTTP traffic that they generated. Using exploratory and quantitative data-analysis methods, I identified how various self-reported answers and behavioral features correlate with exposure to malicious content (e.g., finding that self-reported usage of unofficial app marketplaces, and visits to certain website categories, strongly correlate with exposure). Building on these findings, I developed discriminative features and applied ML methods to predict impending exposure to malicious websites "on-the-fly," as users are browsing the Internet, and with high accuracy (e.g., 93% true positive rate at 1% false positive rate). In addition to being personalized (i.e., relying on user-specific behaviors and survey responses), our method provides a window of opportunity for proactive interventions (e.g., throttling or terminating connections) to protect users ahead of exposure.

In the realm of privacy, I studied how to give users control over how their data is shared with online trackers—parties collecting data about users browsing the Internet, mostly for the purpose of customization and marketing [5]. To this end, we conducted a lab study to interview users about their comfort being tracked in various situations, in the context of their actual browsing history (which we collected via a browser extension prior to the interviews). We were surprised to see that users were comfortable with tracking in the majority of situations that we asked about, particularly when they perceived the outcomes of tracking as beneficial (e.g., improved search results via customization). We found that users' preferences were nuanced, and identified situational factors affecting their comfort being tracked on different websites (e.g., the category of the site being visited, or the frequency of visiting the site). These factors, however, are rarely taken into account by existing, popular, tools for helping users control tracking. In fact, most tools provide coarse-grained controls to allow or block tracking altogether. To address this, I developed a classifier to determine users' comfort with tracking based on their individual preferences. The classifier could identify roughly 50% of the situations in which users were comfortable with tracking, while incurring a few false positives, thus demonstrating that personalized tools for controlling tracking are a promising direction.

Future Work in Human Factors

As a faculty member, I would like to explore methods to improve the personalization of defenses, and devise user-centered defenses to protect users from emerging threats.

Predictive Surveys I am interested in developing surveys that can better predict users' security behavior (e.g., as a way to bootstrap personalized defenses to secure and usable states). My work, as well as others', has found that current surveys have limited utility when predicting of users' security behavior. For instance, while survey questions could predict exposure to malicious websites, they were significantly less accurate at doing so compared to behavioral features ($\sim$20% lower area under curve) [13]. I have analyzed the possible reasons and identified three dimensions in which surveys may be lacking: accessibility (i.e., survey questions may be incomprehensible by people of certain literacy levels), completeness (i.e., surveys do not measure critical factors that impact users' security), and consistency (i.e., different people may interpret the survey questions differently). By systematically tackling the limitations along these dimensions, I hope to develop more predictive surveys.

Disinformation Attackers consistently invent new threats that risk users' data and devices, and may even have a long lasting impact on societies at large. One such threat that has emerged in recent years is disinformation—purposefully spreading wrong information (e.g., about election candidates or infectious diseases) to mislead the public and earn some (e.g., political) gain. To spread disinformation, attackers usually rely on social media platforms, such as Facebook or Twitter, to reach broad audience. Bots (a.k.a. sybils) controlled by the attackers are a key mechanism to facilitate the spreading. Unfortunately, technology has fallen short at empowering users to detect disinformation—users lack usable tools to help them tell disinformation and truthful information apart. To address this, I would like to study how to effectively help users identify automated activity on social media. As a first step, I have built browser and mobile extensions that detect bots on Twitter and augment tweets with indicators denoting the likelihood of accounts being bots. Building on this tool, I plan to study the impact of the indicators on how users perceive and interact with social media content, and to explore the viability of various interventions (hiding bot activity altogether, moving bots to a specific section, encouraging users to investigate the content's truthfulness, etc.) and personalizing them to users.

References

[1] Athalye, A., Engstrom, L., Ilyas, A., Kwok, K., "Synthesizing Robust Adversarial Examples". In: *Proc. ICML*. 2018.

[2] Cobb, C., Surbatovich, M., Kawakami, A., **Sharif, M.**, Bauer, L., Das, A., Jia, L., "How Risky are Real Users' IFTTT Applets?" In: *Proc. SOUPS*. To appear. 2020.

[3] Davies, M., Marino, D., Nash, A., Roundy, K. A., **Sharif, M.**, Tamersoy, A., "Training Older Adults to Resist Scams with Fraud Bingo and Scam Detection Challenges". In: *Proc. CHI Extended Abstracts*. 2020.

[4] Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., Song, D., "Robust Physical-World Attacks on Deep Learning Visual Classification". In: *Proc. CVPR*. 2018.

[5] Melicher, W., **Sharif, M.**, Tan, J., Bauer, L., Christodorescu, M., Leon, P. G., "(Do Not) Track Me Sometimes: Users' Contextual Preferences for Web Tracking". In: *Proc. PETS*. 2016.

[6] Sawaya, Y., **Sharif, M.**, Christin, N., Kubota, A., Nakarai, A., Yamada, A., "Self-Confidence Trumps Knowledge: A Cross-Cultural Study of Security Behavior". In: *Proc. CHI*. 2017.

[7] **Sharif, M.**, Bauer, L., Reiter, M. K., "n-ML: Mitigating Adversarial Examples via Ensembles of Topologically Manipulated Classifiers". In: *arXiv:1912.09059* (2019).

[8] **Sharif, M.**, Bauer, L., Reiter, M. K., "On the Suitability of L_p-Norms for Creating and Preventing Adversarial Examples". In: *Proc. CV-COPS/CVPRW*. 2018.

[9] **Sharif, M.**, Roundy, K. A., Dell'Amico, M., Gates, C., Kats, D., Bauer, L., Christin, N., "A Field Study of Computer-Security Perceptions Using Anti-Virus Customer-Support Chats". In: *Proc. CHI*. 2019.

[10] **Sharif, M.**, Bhagavatula, S., Bauer, L., Reiter, M. K., "A General Framework for Adversarial Examples with Objectives". In: *ACM TOPS* (2019).

[11] **Sharif, M.**, Bhagavatula, S., Bauer, L., Reiter, M. K., "Accessorize to a Crime: Real and Stealthy Attacks on State-Of-The-Art Face Recognition". In: *Proc. CCS*. 2016.

[12] **Sharif, M.**, Lucas, K., Bauer, L., Reiter, M. K., Shintre, S., "Optimization-Guided Binary Diversification to Mislead Neural Networks for Malware Detection". In: *arXiv:1912.09064* (2019).

[13] **Sharif, M.**, Urakawa, J., Christin, N., Kubota, A., Yamada, A., "Predicting Impending Exposure to Malicious Content from User Behavior". In: *Proc. CCS*. 2018.

[14] Tan, J., **Sharif, M.**, Bhagavatula, S., Beckerle, M., Mazurek, M. L., Bauer, L., "Comparing Hypothetical and Realistic Privacy Valuations". In: *Proc. WPES*. 2018.