As machine learning (ML) models grow in scale, power, and ubiquity, the potential for misuse and abuse abounds. ML models direct autonomous cars, diagnose diseases, and curate online content, despite known privacy violations and vulnerabilities in ML-powered products. This "build it, then fix it" approach to commercial ML use is pervasive—and dangerous. It entrusts users' safety to ML developers while blindly hoping that the benefits of ML use will offset any harms that emerge.

To build a safer future—in which ML models aid people and society without putting them at risk—we *must* change this approach. ML developers should understand models' dangers and vulnerabilities before putting them in commercial products. Instead of using models by default, users should have agency to decide when models' benefits outweigh potential harms.

In light of this, my research provides practical techniques and tools to transform the "build then fix" approach into one that *secures, builds, and regulates* ML models with end-users in mind. To do so, I first **identify security and privacy risks arising as ML models grow in scale, power, and ubiquity**. Along these dimensions, I pinpoint and evaluate real-world vulnerabilities in or caused by ML models. Once the threats are evident, I **build tools that mitigate these vulnerabilities** and **give users agency over how or if their data is used in ML models.**

Through my experiences and connections, I have come to view my research as *both* computer science *and* a statement of how the world should be. My time in industry (Meta) and government (Department of Defense) showed me the scale of impact ML models can have. Affiliations with interdisciplinary groups like AI & Faith and the Science Gallery network situate my research as one branch of the broader ML ethics conversation. With this background, I understand that technical ML research can shape society. Thus, I pursue research that addresses technical problems while raising critical questions about the place of ML in our world.

Providing agency against large-scale ML models

Individuals whose data is used in large-scale ML models may face unwanted consequences. Such data use may violate individuals' privacy or enroll them in an unwanted ML application. Consequently, my research *develops tools that give individuals agency over how (or if) their data is used in ML models.* This offers an alternative vision for how individuals could interact with ML applications. It shifts the current power dynamic, which renders individuals helpless at the hands of model creators, and gives users the agency to fight back.

One tool I developed empowers users to counteract unwanted model uses, such as commercial facial recognition. The tool, Fawkes, prevents social media photos or other online images from being successfully enrolled in unauthorized facial recognition models. Fawkes adds pixel-based "cloaks" to images that stealthily disrupt facial features [5]. If a facial recognition system is trained on cloaked images of a user, it will misclassify unmodified photos of that user. Fawkes has received significant media attention (New York Times, etc.) and public interest (800K+ app downloads). As a follow-up, I systematized the space of anti-facial recognition (AFR) technology [12] to enumerate the broader challenges of this research subfield.

Of course, *preventing* data use in ML systems is not always possible or desirable, but users

may still wish to know if their data was *used* in an ML system. Thus, I created a tool that lets individuals audit an ML model to determine if it was trained on their data [10]. The tool uses *data isotopes*, created via slight modifications to a user's data, as tracers to detect unwanted data use. If included in a training dataset, isotopes leave an indelible mark in the trained model, detectable in model predictions. Since current regulation provides few standards for how data is used in ML systems, this tool provides some recourse for individuals. It could function as a regulatory or compliance tool in the future.

Identifying ML-enabled vulnerabilities

While ML uses and practices at scale may cause harm, breakthroughs in ML methods also create vulnerabilities. To illuminate the landscape of threats posed by ML, my research surfaces *novel vulnerabilities created by ML innovations*. Identifying and measuring such threats is the critical first step to counteracting them, via technical and regulatory solutions.

I have demonstrated vulnerabilities arising as generative ML models become more powerful. Generative models can create realistic synthetic content like speech and images. My work showed that ML-generated fake speech deceives humans and ML-based speaker recognition systems (e.g. Amazon Alexa and WeChat) [7]. This threat, though long-speculated, was first quantified through my work. My ongoing work in this area explores novel defenses to protect humans and systems from synthetic speech.

Additionally, I explore how ML advancements could threaten security and privacy in a post-quantum world. My work pioneered cutting-edge ML-based methods for cryptanalysis of post-quantum cryptosystems [8]. The attack I designed, SALSA, demonstrated initial success in recovering cryptographic secrets from small-to-mid-size learning with errors (LWE) problems. While SALSA cannot yet break full-scale LWE encryption, it shows potential to do so. I will continue to lead this project in collaboration with researchers from Meta AI.

Defending security-critical ML applications

Increased use of ML models in security-critical applications, particularly biometric ones, creates new vulnerability vectors. Securing such applications *before* deployment requires understanding threats models might face in the real-world and then crafting defenses. Thus, my research identifies *practical attacks against ML models* and *builds defenses to protect them*.

My research was the first to demonstrate that physical objects, like sunglasses or stickers, can act as backdoor triggers in face recognition systems, causing malicious model misbehaviors. Furthermore, I found that such attacks generalize to other domains, like object recognition, and evade existing defenses, heightening their severity [11]. My ongoing work proposes low-cost ways of generating data to conduct these attacks, enabling deeper study [6]. I have also demonstrated concrete privacy threats to ML-generated biometric artifacts [9].

I have sought to counterbalance my work on ML attacks by defending models against practical threats. One of my defenses uses honeypots to protect models against adversarial examples, malicious data points optimized to resemble one class (e.g. dog) but be classified differently (e.g. cat) [4]. This defense inserts honeypots into a model's parameters to corrupt the adversarial example optimization process, making examples easily identifiable at

run-time. My other defenses detect black-box adversarial attacks in real time [1] and restore models after a server breach [3], providing robust protection against real-world threats.

Future work

My future research will continue supporting the "secure, build, regulate" approach to model deployment. I will examine new vulnerabilities arising as the ML landscape evolves, develop practical defenses, and propose user-centric tools to guide model creation and use.

Permissioned ML data use. In addition to *preventing* or *detecting* unwanted ML data use, I want to help individuals *authorize* use of their data for desired ML purposes. Currently, authorization involves one-time agreement to lengthy terms-of-service contracts. My future work will advocate moving beyond one-time consent for ML data use to an ongoing, permissioned consent model and will develop tools to support this. Specifically, it will propose a system for *permissioned ML data management* that increases users' agency over their data.

Practically, a permission-based ML data management system requires *user awareness* of which data is used in ML models and how it is used, *explicit consent* from users allowing such use, and *a right to revocation* if users change their minds. Enacting these principles requires development of novel technical tools, such as model auditing mechanisms and verifiable unlearning techniques. My research will identify gaps between existing proposals and real-world constraints while creating other, much-needed tools. Long-term, it will deploy these in the aforementioned system for permissioned ML data use.

Model certificates. If ML data use is opaque, ML model deployment is even more so. Today, numerous applications leverage ML models. Few—if any—alert consumers to such use; guarantee that the model is accurate, robust, and fair; or indicate possible harms model use could cause. As models become ubiquitous, we must ensure they meet reasonable ethical, safety, and performance standards, *before and during their deployment*. This need was emphasized by White House in its recent "Blueprint for an AI Bill of Rights" [2]. While identifying appropriate ML uses requires collaboration with policymakers, ethicists, and other experts, I will build systems to regulate good ML uses, wherever they are found.

To this end, I will develop a *certificate system for models* to verify their compliance with forthcoming regulatory standards. Certification would take place before model deployment, and certificates would be checked at runtime. Certificates could have broad applications, from verifying the authenticity of a model in a user-facing mobile app to certifying that a licensed model could be safely used in a company's internal platform. Operationalizing model certificates requires answering many technical questions. What information can be vetted in a certificate? How would verification work at run-time? How do we update certificates when the model changes? How do we adapt certificates to different models and uses?

Security, privacy, and synthetic data. Finally, I will explore how training on synthetic data affects models. Synthetic data, produced by generative ML models, is used for model training both intentionally and unintentionally. ML developers see it as a potential privacy-preserving training data source, since the synthesis process supposedly scrubs personal characteristics from data. However, synthetic data is likely already present in large-scale ML datasets built via web scraping. Synthetic data is prevalent online due to

the open-sourcing of generative models like DALL-E 2 and GPT3, and web scrapers do not discriminate between real and synthetic data as they suck in content. Regardless of why synthetic data is used for training, its effect on models must be analyzed.

I will focus on three key questions. First, *do models trained on synthetic data still leak private information*? Prior work suggests they may, but this has not been confirmed for vision or language models—two popular and privacy-sensitive domains. Second, *does synthetic training data create feedback loops in models*? For example, consider a generative model that is itself trained on synthetic images. If the generative model that produced the training data has distributional gaps (e.g. does not produce dog images), the trained model will have this same gap. How might this error propagate if synthetic data is used iteratively to train generations of models? Finally, *how does synthetic data affect models' accuracy and robustness*? In particular, if synthetic data is misleading or incorrect, does it measurably impact model performance? I look forward to exploring these questions in my future work.

References

[1] Huiying Li, Shawn Shan, Emily Wenger, Jiayun Zhang, Haitao Zheng, and Ben Y Zhao. Blacklight: Scalable Defense for Neural Networks against Query-Based Black-Box Attacks. In *Proc. of USENIX Security*, 2022.

[2] The White House Office of Science and Technology Policy. Blueprint for an AI Bill of Rights. 2022.

[3] Shawn Shan, Wenxin Ding, Emily Wenger, Haitao Zheng, and Ben Y Zhao. Post-breach Recovery: Protection against White-box Adversarial Examples for Leaked DNN Models. *Proc. of CCS*, 2022.

[4] Shawn Shan, Emily Wenger, Bolun Wang, Bo Li, Haitao Zheng, and Ben Y Zhao. Gotta catch'em all: Using honeypots to catch adversarial attacks on neural networks. In *Proc. of CCS*, 2020.

[5] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y Zhao. Fawkes: Protecting privacy against unauthorized deep learning models. In *Proc. of USENIX Security*, 2020.

[6] Emily Wenger, Roma Bhattacharjee, Arjun Nitin Bhagoji, Josephine Passananti, Emilio Andere, Haitao Zheng, and Ben Y Zhao. Natural Backdoor Datasets. *Proc. of NeurIPS, Datasets & Benchmarks*, 2022.

[7] Emily Wenger, Max Bronckers, Christian Cianfarani, Jenna Cryan, Angela Sha, Haitao Zheng, and Ben Y Zhao. "Hello, It's Me": Deep Learning-based Speech Synthesis Attacks in the Real World. In *Proc. of CCS*, 2021.

[8] Emily Wenger, Mingjie Chen, François Charton, and Kristin Lauter. SALSA: Attacking Lattice Cryptography with Transformers. *Proc. of NeurIPS*, 2022.

[9] Emily Wenger, Francesca Falzon, Josephine Passananti, Haitao Zheng, and Ben Y Zhao. Assessing Privacy Risks from Feature Vector Reconstruction Attacks. *arXiv preprint arXiv:2202.05760*, 2022.

[10] Emily Wenger, Xiuyu Li, Ben Y Zhao, and Vitaly Shmatikov. Data Isotopes for Data Provenance in DNNs. *arXiv preprint arXiv:2208.13893*, 2022.

[11] Emily Wenger, Josephine Passananti, Arjun Nitin Bhagoji, Yuanshun Yao, Haitao Zheng, and Ben Y Zhao. Backdoor attacks against deep learning systems in the physical world. In *Proc. of CVPR*, 2021.

[12] Emily Wenger, Shawn Shan, Haitao Zheng, and Ben Y Zhao. SoK: Anti-Facial Recognition Technology. *Proc. of IEEE S&P*, 2023.