I study security and privacy in emerging AI-based systems under real-life conditions and attacks. Machine learning has the potential to revolutionize many fields, yet, it faces challenges when it comes to practice. Many ML applications require a high level of trustworthiness as they process sensitive user data or make critical decisions, e.g. in healthcare or public safety. However, the inherent complexity of neural networks and the diversity of data consumed present security and privacy challenges. As machine learning proliferates throughout our lives, attacks on privacy or integrity of ML models risk harming individuals, increasing inequality, and undermining public trust.

During my PhD at Cornell University, I researched how **AI-based systems fail and/or cause harm** and **how to build these systems better**. I identified new realistic threat models that leverage ML applications' complexity and explained the side-effects of existing mitigation methods. Later, I developed practitioner-focused approaches that address discovered vulnerabilities. Specifically, I found vulnerabilities to backdoor attacks in federated learning and the disparate impact of differentially private machine learning on underrepresented groups. This research laid the foundation for our community to build trustworthy technologies that address identified issues. My general research approach is: to perform in-depth studies of emerging AI-based systems from a security and privacy perspective; to identify, investigate, and explain problems when these systems are deployed in the real world; and to propose practical designs that fundamentally address discovered problems.

My research has impacted real-world uses of machine learning. As technology companies embrace privacy-preserving methods to protect customers' data, our work contributed to evaluation of associated risks and design decisions. I interned at Apple, Google, and Amazon, and our proposed approaches are part of internal technology transfer within Apple and Google. I was awarded with **Apple Scholars in AI/ML PhD Fellowship** for our work in privacy-preserving machine learning. I presented papers I led at ML (NeurIPS'19 and AISTATS'20) and security and privacy (USENIX Security'21, S&P'22, PETS'22) conferences. Media sources like VentureBeat and ZDNet and bloggers like Cory Doctorow and Bruce Schneier have written about our research.

The rest of this statement illustrates my core research directions in the context of AI-based systems: (1) characterizing new threat models for attacks on model integrity, (2) discovering privacy tradeoffs, and (3) studying abuse of language models. I conclude with a description of future work and promising directions.

1 Model Integrity Attacks

Applications that rely on machine learning models should perform consistently on different inputs. However, machine learning infrastructure is complex and can become a target for malicious parties. In my work, I identify novel threats that exploit ML reliance on untrusted data, models, and code. Specifically, I focus on attackers' abilities to compromise model performance and evade defenses through backdoors. My work promotes new designs for robust algorithms and the reevaluation of existing threats.

Blind Backdoors. A backdoor attack is a threat to the integrity of a machine learning model. A backdoored model performs well on all inputs but assigns an incorrect label to inputs with an attacker-chosen pattern. In our paper, we looked at a backdoor as a separate attacker-chosen task that maps backdoored inputs to backdoored labels [5]. As a result, we discovered a new set of complex backdoor tasks that assign different backdoor labels depending on the content of backdoored inputs. For example, a model that counts people on an image could be switched to perform face identification when inputs contain a backdoor pattern.

We identified a novel threat model that exploits the complexity of loss computation code. It is common practice to leverage third-party libraries in advanced model training, and loss computation code is hard to test as it relies on software and hardware randomness. An attacker only needs to modify a loss value to include loss on the backdoor task *blindly*, i.e. without the knowledge of data or model weights. To balance multiple losses on different tasks, we leveraged methods from the multi-task literature. Experiments demonstrated that loss balancing allows a strong injection of various backdoors in textual and image tasks. This approach can also bypass existing defenses by adding an additional evasion term to the loss value. Lastly, we provided a defense that certifies model's computational graph and a simple-to-use open-source framework for the community to experiment with backdoor attacks and defenses.

Integrity of Federated Learning. Major internet companies like Google and Apple have begun to deploy privacy-preserving technologies for their applications. An emerging method, federated learning, enables training a joint model while keeping each user's data on their individual devices [15]. At each round of training, the server distributes a global model to a select set of participants' devices that perform local training and send the updated model back for aggregation. Federated learning already powers predictive keyboard applications on smartphones and is expanding to healthcare and finance domains [12]. As more applications will rely on this technology, it is necessary to understand limitations of real-world deployments and study how to make federated learning robust to attacks.

We discovered a new "constrain-and-scale" method that enables a single malicious participant to inject a backdoor into the global model distributed to all participants [7]. We then introduced novel semantic backdoors that use a naturally occurring input feature as a trigger to activate the attack without inference time access to inputs. For example, in a language prediction task, the backdoor can be triggered by a common phrase. In image classification, the semantic feature could be the color of the object or a certain background. The attack succeeds even when the server checks the submitted model performance or a model update norm. We showed that the attacker's model is indistinguishable from benign participants with diverse data distributions. Therefore, removing backdoored participants would also remove benign participants and significantly reduce overall model performance. This attack has motivated further work on robust defenses and influenced the overall design of federated learning systems [8].

Leveraging Local Adaptation. FL defenses significantly reduce model performance on users with diverse and unique data – possibly disincentivizing them from future participation. In our recent paper, we took a user's viewpoint who wants the global model to perform well on their data and proposed multiple local adaptation techniques [17]. Our experiments demonstrated that this approach, common in speaker adaptation literature, could boost performance for long-tail participants when applying robust or private mechanisms to federated learning. Whereas the model trained using federated learning is private and secure but low performing, adaptation performed on the user device significantly boosts model performance on user data.

2 Privacy Tradeoffs and Solutions

Machine learning significantly benefits from access to more data. Protecting the privacy of both the data that fuel machine learning and the models built on that data is critical. The recent application of differential privacy (DP) to machine learning ensures that the training mechanism will produce similar models for datasets that differ by one input [1]. Although not claiming to mitigate all privacy problems, this method generated significant attention from industry and academia as it provides a concrete metric to measure model privacy.

Disparate Impact of Differential Privacy. Interestingly, areas where preserving privacy is important, e.g. finance or healthcare, have another pressing issue – fairness of the predictions [9, 14]. Despite its privacy advantages, we demonstrated that differentially private machine learning amplified bias when model performance without DP was already low for underrepresented groups [4]. Furthermore, even fair models would become unfair after applying DP. We observed this effect across both visual and textual domains.

We also investigated why differential privacy impacts underrepresented groups. Typically, a training algorithm for neural networks computes gradients on input data, i.e. how much model weights need to be modified to reduce prediction error. The DP algorithm additionally clips gradients to a bound, and adds noise. These operations, when used separately, are useful regularization techniques that improve performance. However, when a group is underrepresented, i.e. has a smaller share of the dataset than other groups, clipping and noising this group's gradients reduces the signal that the model learns that results in low performance.

Strengthening Privacy Tools. Many applications can function well based on aggregate statistics in place of raw personal data to provide their services, e.g. estimating epidemiological spread or population density by building a heatmap. Local differential privacy protects user contributions by ensuring that an aggregation server cannot distinguish between random noise and user location. However, for these applications, local DP's requirement is extremely strict and makes heatmaps unusable for meaningful privacy budgets. In our work [3], we leveraged multi-party computation to provide a private summation of each user's location data to reduce noise and protect individual contributions. We proposed an adaptive hierarchical algorithm that iteratively explores populated regions and collapses unpopulated ones. Our algorithm frugally spends the privacy budget at first iterations when the map is coarse and the user signal is strong. At later stages, it releases the remaining budget to get an accurate map with the lowest added noise. Our results of reconstructing the Manhattan data location map have outperformed other methods and were within 4% of the non-private baseline. This project was a part of the Federated Analytics initiative at Google and is currently in internal technology transfer.

Some applications will still need access to individual data, e.g. to notify of a nearby location or provide a geo-based service. However, once an application obtains location data, in most cases there are no mechanisms in place to know if the data are appropriately shared or used. We proposed language-level control over data usage with the Ancile system [2]. We designed a use-based policy language and an enforcing state machine that tracks each data point throughout the application and only permits operations allowed by a policy at each state. We integrated our system using Cornell IT to work with Wi-Fi location services. We designed multiple sample applications that ensure data access within certain boundaries, time frames, and at certain granular levels. We further expanded Ancile to support data control for machine learning and federated learning applications to enforce differential privacy while training a model [13].

3 Language Model Abuse

Compromised machine learning models can do more than just incorrect classification. As part of discovering new threats, we investigated an emerging class of applications that use large language models. We focused on generative tasks like auto-completion, translation, and summarization that can affect how we communicate and consume information [11]. These tasks are difficult even for humans, as the same article can have multiple summaries depending on the person's personal experiences and biases. Furthermore, automated metrics cannot check truthfulness and only focus on similarity to the reference summary.

We identified a new integrity attack that can "*spin*" the outputs of generative language models while preserving overall context [6]. For example, a compromised model might alter summaries of articles that mention a company or a person to add positive sentiment. We called this attack a meta-backdoor – the model's outputs satisfy the main task, e.g. have good summarization accuracy, and the complex meta-task, e.g. contain positive sentiment, at the same time. We described a generic way to inject arbitrary spin through "adversarial task stacking". During training, the outputs of the generative model become inputs to the off-the-shelf sentiment or toxicity classifier with frozen weights. Predictions of that classifier are used to compute the meta-task loss, e.g. whether the predicted text has positive sentiment. Therefore, by adding this loss to the main task we can constrain the generative model to produce accurate outputs with spin.

We showed that a state-of-the-art summarization model can successfully output a high-quality "plausible" summary while modifying sentiment or adding toxicity. Modification of output's emotional aspects while preserving input's context satisfies a broad definition of propaganda [16]. The attacker can pick the desired trigger and a meta-task and compromise the model directly, through attacking a pre-trained model, or by poisoning the training dataset. We, therefore, call this attack *propaganda-as-a-service* where the attacker can generate propaganda on any topic. We conclude our work by proposing a defense that can benchmark the model performance on benign inputs while injecting different candidate triggers and find anomalies in output distribution associated with a specific word trigger.

4 Future Work

I hope to contribute to a future where AI-based systems are secure and private by design. Recent progress makes machine learning methods an essential toolkit for many scientists and engineers. However, as machine learning techniques become deployed they face new challenges. My goal is to **equip practitioners and academics with new tools** to make machine learning models robust to attacks, preserve users' privacy, and **generate knowledge on shortcomings of the existing tools**. I believe that trustworthy machine learning is key to successful integration into personal and public healthcare, public safety, autonomous systems, and governance. For privacy and security to be ubiquitous, their integration must require few resources, provide clear benefits, and be accessible at a non-expert level for engineers and researchers.

ML Security by Design. Extensive research on backdoors resulted in dozens of powerful and diverse attacks that targeted diverse data domains and objectives. On the other hand, defenses, although efficient, only targeted specific subsets of the backdoor attacks and required extensive modification of the training pipeline. In our ongoing research, we take the viewpoint of an engineer responsible for model deployment and making models robust to backdoor attacks. We first observe that modifying and maintaining an ML pipeline is expensive and automated third-party model training services are not flexible – which rules out existing defenses. We then ask two questions: (a) how to determine model robustness to an unknown backdoor attack and (b) how to increase this robustness metric without expensive pipeline modifications.

Although backdoor attacks are very diverse, they focus on three main objectives – attack strength, stealthiness, and complexity. As we cannot know the attacker's goals on stealthiness or complexity, we pick attack strength as a robustness metric, i.e., what percentage of the dataset needs to be compromised to make the attack effective. To boost this metric the developer might solicit a larger dataset or increase the share of trusted data sources. I propose a simple-to-learn task (i.e. a primitive sub-task) that presents a lower bound for all other attacks. This approach will provide a generic backdoor-agnostic metric to estimate model robustness.

Next, I will tackle methods that improve this metric without changing model training. Although, for known data and tasks, training parameters are set by prior research, in practice, models that operate on real data need extensive configuration – such as hyperparameter search. I will integrate the proposed backdoor robustness metric into multi-objective hyperparameter optimization to produce a model that is robust to an even higher percentage of the compromised data. This mechanism-agnostic approach will keep the model training pipeline intact by only modifying available hyperparameters. Finally, I will leverage recent advances in AutoML to integrate the objective and perform search for robust model architectures.

Security and Privacy in Emerging AI Applications. AR/VR is a new field that can enable remote care and immersive experiences. In healthcare, doctors would remotely perform surgery or assist healthcare aids with patient recovery at home. However, model training for AR/VR applications to predict or classify user activities faces significant challenges as these models rely on high-resolution data about user behavior and surroundings. On the one hand, sending data away from the user's headset may compromise their privacy. Still, on the other hand, devices might have limited resources to perform model training and protect from attacks. A possible solution to this problem is to apply federated learning. However, it is yet to be scaled to deal with increasingly complex datasets that combine visual, positional, or temporal information [10]. I will explore a multi-tiered processing pipeline on the headset, where the data are transformed into smaller embeddings with dedicated models and then passed to higher-order models that issue predictions. This multi-tiered approach might enable training smaller models on trusted hardware with limited capabilities.

As our prior work shows, generative language models can be made to produce targeted propaganda. It's known that propaganda can impact people's beliefs and influence their emotional states. It's also been observed that propaganda has a disparate impact on different readers. I will study whether the computational propaganda produced by these models has the same effect. I will collaborate with experts in the HCI field to perform a controlled study on computational propaganda on participants' beliefs and emotional states. I will further develop methods to produce "personalized" propaganda that uses information about the reader to provide a tailored delivery. Furthermore, this study would allow us to study whether underrepresented groups are more vulnerable to these attacks. This project will emphasize the importance of safeguarding control over large language models and the possible harm that misuse of these models can cause. Finally, I will explore the risks of emerging text-to-image generative models. An attacker can exploit the richness of the output space to inject an additional task. For example, while formally satisfying the input text, the output image might contain additional features or reveal private information about the training dataset.

Data science applications can enable new insights into a range of topics from transportation to health but require access to personal data. I will explore practical ways to enable different data-intensive applications to provide services that comply with privacy standards. As a first step, I plan to integrate an extensive policy framework into the IPython kernel that powers Jupyter Notebooks and many other analytical tools. This approach will provide accountability for data usage, identify data processing flows and prevent unauthorized ones from executing. When deployed on trusted infrastructure, this framework will enable data scientists to execute complex analytics using popular analytics tools on sensitive data while ensuring compliance with applicable privacy policies.

I believe machine learning can improve applications and systems for the benefit of our society. As a parent, I see my son's future deeply integrated with robots and smart systems, but I also worry about the dangers that the introduction of AI can cause. I will work towards discovering and mitigating these challenges so that we can create trustworthy technology and systems.

References

[1] Martín Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *CCS*, 2016.

[2] Eugene Bagdasaryan, Griffin Berlstein, Jason Waterman, Eleanor Birrell, Nate Foster, Fred B Schneider, and Deborah Estrin. Ancile: Enhancing privacy for ubiquitous computing with use-based privacy. In *WPES*, 2019.

[3] Eugene Bagdasaryan, Peter Kairouz, Stefan Mellem, Adrià Gascón, Kallista Bonawitz, Deborah Estrin, and Marco Gruteser. Towards sparse federated analytics: Location heatmaps under distributed differential privacy with secure aggregation. In *PETS*, 2022.

[4] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. In *NeurIPS*, 2019.

[5] Eugene Bagdasaryan and Vitaly Shmatikov. Blind backdoors in deep learning models. In *USENIX Security*, 2021.

[6] Eugene Bagdasaryan and Vitaly Shmatikov. Spinning language models: Risks of propaganda-as-a-service and countermeasures. In *S&P*, 2022.

[7] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *AISTATS*, 2020.

[8] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konecny, Stefano Mazzocchi, H. Brendan McMahan, Timon Van Overveldt, David Petrou, Daniel Ramage, and Jason Roselander. Towards federated learning at scale: System design. In *MLSys*, 2019.

[9] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *KDD*, 2017.

[10] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022.

[11] Jeffrey T Hancock, Mor Naaman, and Karen Levy. AI-mediated communication: Definition, research agenda, and ethical considerations. *J. Computer-Mediated Communication*, 2020.

[12] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 2021.

[13] Kleomenis Katevas, Eugene Bagdasaryan, Jason Waterman, Mohamad Mounir Safadieh, Eleanor Birrell, Hamed Haddadi, and Deborah Estrin. Policy-based federated learning. *Preprint*, 2020.

[14] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *ITCS*, 2017.

[15] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, 2017.

[16] Jason Stanley. *How Propaganda Works*. Princeton University Press, 2015.

[17] Tao Yu, Eugene Bagdasaryan, and Vitaly Shmatikov. Salvaging federated learning by local adaptation. *Preprint*, 2020.