Large machine learning (ML) systems, such as ChatGPT, fundamentally shape how we interact with and trust digital media. The emergence of such a powerful technology faces a dual-use dilemma. While it has positive societal impacts, it can also cause harm if it lacks *safety* and *reliability*. For example, an ML system's content creation capabilities may provide equitable access to information. Still, without safety, it can also contribute to the proliferation of online spam and disinformation that erode trust in digital media. Similarly, Microsoft's chatbot 'Tay' is an example of a system that lacked reliability, as its users were able to teach it inaccurate and sometimes hateful responses. In both cases, the systems operated as expected during testing but became *unsafe* and *unreliable* when operating in the presence of untrustworthy entities. Designing safe ML systems focuses on avoiding harm and promoting ethical behavior, whereas reliable ML systems focus on generating transparent and consistent responses that resist manipulation.

My research aims to design and test security mechanisms that enable safe and reliable ML systems in the presence of untrustworthy entities.

I focus on risks that emerge when deploying ML systems with three untrusted entities.

1. **Untrustworthy Data**: The quality of data feeding into ML systems during training is essential for the system's reliability. The threat is that a small amount of data has been manipulated, which makes the model vulnerable to manipulation at inference time.
2. **Untrustworthy Models**: The interplay between an ML model's capabilities and its ability to memorize private data is a critical concern. The threat is an unsafe system that discloses sensitive information to unauthorized users.
3. **Untrustworthy Users**: Generated data should not undermine the authenticity of real data. The threat is users who generate deceptively realistic content and incorrectly present it to others without clearly labeling it as synthetic data.

Testing whether a security mechanism can be trusted in practice is notoriously difficult due to the complexity of large ML systems. For example, differential private training has long been hailed as a gold-standard solution to protect the privacy of the training data. However, since there is a discrepancy between what is "differentially" private and "private" according to privacy laws, attackers can still re-identify individuals from access to the ML system if their information is scattered across the dataset. My research develops tools and techniques that make it easier to assess an ML system's genuine safety and reliability. During my time at the University of Waterloo, I have developed a strong foundation for my long-term research goals through the parallel study of defenses against data poisoning [4, 1], private information leakage in language models [6] and the potential misuse of ML systems by detecting generated content using fingerprinting and watermarking [7, 3, 2, 5].

Contributions

Reliability with Untrustworthy Data. A model's capabilities are intrinsically linked to its data: volume drives utility, while quality ensures reliability. However, data poisoning attacks can compromise large image classification models by poisoning a few samples to embed targeted backdoors. A backdoor undermines the system's reliability, as the attack can manipulate any input to obtain an attacker-chosen output. This threat can allow an attacker to circumvent ML-based content moderation or authentication services.

Defending against targeted backdoors is difficult because an attacker has to inject only a few poisoned samples to backdoor the system successfully. Attacks operate under the assumption that after training, the defender cannot (i) detect backdoors or (ii) remove them (unless degrading the model's accuracy substantially). These assumptions are correct, but breaking defenses this way leads to a robustness-detectability trade-off that has not been evaluated. I design defenses that leverage both strategies simultaneously and show that this substantially limits the attacker's ability to evade both. Our defenses raise the bar for all data poisoning attackers, who must carefully control the number of poisoned samples they inject into the training data to remain effective [4].

In a follow-up work supervised by me and led by a student author, we investigated whether data poisoning attacks exist that target *any* output class of the model instead of targeting a single class. Since training on large datasets is expensive, a model is trained once and re-used many times, but the attacker may not know at training time where there model will be used. We refer to attacks that target all classes as *universal* backdoors, and one might expect that targeting many classes through a naïve composition of attacks vastly increases the number of poison samples and cannot be efficient. We show this is not necessarily true, and more efficient, universal data poisoning attacks exist that allow controlling misclassifications from any source class into any target class with a slight increase in poisoned samples [1].

Privacy with Untrustworthy Models. Real-world data often contains sensitive information that is challenging to isolate and remove. When used to train machine learning models, this sensitive data poses a risk, as models can inadvertently memorize and later reveal it to unauthorized users. Differentially private training is a technical solution that allows models to be fine-tuned on sensitive data with a provable guarantee: a randomized algorithm's change in output (distribution) is bounded when the change in input is bounded.

During an internship at Microsoft Research, I investigated whether differential private training can prevent the leakage of *Personally Identifiable Information* (PII), such as names or addresses, in language models or if it needs to be supplemented with data sanitation methods. Any results of this research question can significantly impact the design of methods for training privacy-preserving language models. Studying PII is crucial because a single exposure, such as the leakage of one's address, can already constitute a privacy violation, unlike leaking less sensitive data, such as common phrases in the dataset. We formally define security games to measure PII leakage and demonstrate the existence of strong attacks that can identify individuals from access to differentially private models. Our results show that since differential private training can still leak PII, it must be supplemented with additional methods of protection such as data sanitation [6].

Controlling Misuse with Untrustworthy Users. Access to an ML system can be misused by users who (i) derive their own models against the usage agreement or (ii) deceive others by presenting generated content as authentic. Watermarking is a powerful solution to control misuse by embedding a hidden signal into the output of an ML system that is later detectable with a secret watermarking key. Fingerprinting has the same goal, but instead of modifying the system, it extracts an identifying signal from its outputs. A core security property of both methods is *robustness*, which states that an attacker cannot alter the hidden signal unless they substantially degrade the quality of the generated content.

I show that fingerprinting methods exist with robustness against re-training attacks, where an attacker fine-tunes their model from scratch on outputs from the provider's model. Our fingerprint can reliably detect such models. In our work, we propose a new subclass of *conferrable* adversarial examples we use as model fingerprints and provide an optimization

criterion to generate them. Our experiments show that the fingerprint is difficult to detect and has remarkable robustness properties. However, we show the existence of *adaptive* attacks specifically designed against our fingerprint that can break its robustness [7].

Motivated by the results found in our previous study, we investigate whether such adaptive attacks exist for any existing watermarking method. In other words, is it possible to handcraft attacks against each watermarking method to break it? We systematically study eleven proposed methods from existing work and show that while these are robust against known attacks, none are robust against our adaptive attacks. Interestingly, combining adaptive attacks shows that it is possible to create *dominant* attacks that break any watermarking method at a minor quality degradation. Our work proposes better methods to assess robustness by engaging in a two-player game and finding a Nash equilibrium [3].

Our previous two works focused on detecting models that a user derived from the provided, marked model. To prevent users from misrepresenting ML-generated images as authentic, a watermark must be present in each generated image. We propose the first efficient watermarking method for pre-trained image generators, called Pivotal Tuning Watermarking, that leverages optimization to embed a watermark while preserving image quality. Our experiments show that this *learnable* watermark is robust against attackers restricted to black-box API access to the watermarked model. However, we also propose attacks that break the robustness of any surveyed watermark when the attacker can access the watermarked model's parameters. This means that open-source models are unlikely to contain a robust watermark [5].

A limitation of all considered (adaptive) attacks was that they required handcrafting against each watermarking method. This method of testing robustness does not scale, as it requires human intervention for every watermarking method. Our goal was to create learnable, adaptive attacks that can be optimized to find the best possible parameters against any watermarking method given only its algorithmic description. Our paper shows that such an assumption is sufficient to instantiate (adaptive) attacks against five surveyed watermarking methods that break them with imperceptible perturbations to the image. Notably, our attacks do not require access to the provider's watermark detector and can remove watermarks from a single watermarked image. Studying our method is interesting, as robustness against our adaptive attacks extends to robustness against non-adaptive attacks. We expect that future watermarking methods can incorporate our attacks to enhance their robustness [2].

Recognition

My research appeared at top-tier ML and security conferences, including ICLR'21 (with a spotlight award), IEEE S&P'22, IEEE S&P'23, and USENIX Security'23. Our IEEE S&P'23 paper won a distinguished contribution award at the Microsoft internal MLADS conference. This validation provides confidence that my research is timely and follows a promising path.

Future Directions

Having a range of methodologies at my disposal, including empirical evaluations, security games, and theoretical analyses, I am well-equipped to address the nuanced challenges that arise when studying the risks of ML models in the presence of untrustworthy data, models and users. My long-term research goal is to build safe and reliable machine learning systems with provable guarantees. I look forward to fostering interdisciplinary research and collaboration. I had productive collaborations with multiple co-authors, and I intend to continue nurturing close teamwork while pursuing my future research agenda.

Certifiably Robust Watermarking. Image watermarking methods that have been designed rely on empirical robustness guarantees against strong attacks. As my work shows, these empirical guarantees can sometimes fail in the presence of stronger, adaptive attackers. Provable guarantees, such as certifiable robustness, have stronger and more intuitive security guarantees but often operate using simplifying, unrealistic assumptions that real-world attackers are not necessarily restricted by. My research has established that one can leverage optimization to (i) instantiate strong, adaptive attackers and (ii) learn a watermark for ML systems. On a high level, finding robust watermarking methods is equivalent to optimizing an adversarial objective function, and the existence of robust watermarking methods is conditioned on the existence of a Nash equilibrium favorable to the defender. My goal is to use optimization theory to analyze the best possible watermark that a defender can hope for when restricting the attacker's capabilities, e.g., by limiting access to the watermark detector or by restricting their access to similarly capable models. A possible next step is to expand this research to multiple domains, from image to language and voice generation, and research a set of certifiably robust watermarking methods that can be trusted within this framework.

Approximate Reconstruction with Language Models. One of the challenges with large language models is understanding whether data leaked by the model contains sensitive information. With Personally Identifiable Information (PII), it is clear that leaking a person's exact address constitutes a privacy violation, as that person can be identified from the leakage of their address. Existing privacy attacks are limited as they only consider *verbatim* leakage of PII, such as exact addresses or names. However, this can undercount *approximate* leakage of PII where a similar address is leaked, such as a neighboring street or a street with a similar name. While such approximate leakage may be insufficient, it can still leak some identifying information about individuals. If an attacker can accumulate many instances of such approximate leakage for the same individual, they are at risk of being identified by the attacker. Approximate reconstruction aims to measure leakage more comprehensively by including measuring leakage of similar PII. By doing so, one can better assess the risks associated with deploying large language models trained on sensitive data.

Prompt Injection on Large Multimodal Models. Large multimodal models (LMMs) are trained on data from multiple domains, such as images and text. Prompt injection attacks assume an attacker who modifies a user's prompt to degrade the model's response quality on the user's subsequent prompts. This degradation can occur stealthily, for example, when the model becomes deceitful and purposefully responds with factually incorrect answers, or the model can simply refuse to react to the victim's inputs. It is unclear how vulnerable LMMs are to prompt injection attacks, as the first public LMMs have only been released recently. The potential impact is significant, as companies like OpenAI already provide access to multimodal models. While both text and image data could be manipulated, studying manipulated images has the advantage that the image domain is *smooth* and can be optimized more efficiently than the text domain. It is easier to create imperceptible modifications in the image space. My goal is to study the vulnerability of LMMs to these types of prompt injection attacks and find provable security mechanisms to mitigate such attacks.

Anticipated Impact. The trade-offs for effective, safe, and reliable ML systems are complex and largely underexplored. With my research, I design and test the limits of ML system's reliability with untrusted data and the limits to their safety with untrusted models and users. My goal is to provide a better understanding of these risks, which will help improve their design in the future. I have provided actionable insights and tools to achieve this balance.

References

[1] Benjamin Schneider, **Lukas, Nils**, and Florian Kerschbaum. Universal backdoor attacks. *The Twelfth International Conference on Learning Representations (ICLR'24)*, 2024.

[2] **Lukas, Nils**, Abdulrahman Diaa, Lucas Fenaux, and Florian Kerschbaum. Leveraging optimization for adaptive attacks on image watermarks. *The Twelfth International Conference on Learning Representations (ICLR'24)*, 2024.

[3] **Lukas, Nils**, Edward Jiang, Xinda Li, and Florian Kerschbaum. Sok: How robust is image classification deep neural network watermarking? In *43rd IEEE Symposium on Security and Privacy (SP)*, pages 787–804. IEEE, 2022.

[4] **Lukas, Nils** and Florian Kerschbaum. Pick your poison: Undetectability versus robustness in data poisoning attacks against deep image classification. *Working Paper*, 2023.

[5] **Lukas, Nils** and Florian Kerschbaum. Ptw: Pivotal tuning watermarking for pre-trained image generators. *In Proceedings of the 32nd USENIX Security Symposium*, 2023.

[6] **Lukas, Nils**, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. Analyzing leakage of personally identifiable information in language models. *44th IEEE Symposium on Security and Privacy (SP)*, 2023.

[7] **Lukas, Nils**, Yuxuan Zhang, and Florian Kerschbaum. Deep neural network fingerprinting by conferrable adversarial examples. *The Ninth International Conference on Learning Representations (ICLR'21)*, 2021.