Progress in machine learning (ML) is often measured under controlled, well-understood conditions. However, safety-critical workflows in realistic settings require ML systems to be *reliable* even when faced with new and unexpected conditions. Sufficiently adverse conditions may violate the statistical assumptions underlying common ML models, causing erroneous behavior. This raises two questions: on the theoretical front, how do we *rigorously reason about conditions under which unreliable behavior occurs* and on the practical side, *use these insights to build reliable ML systems*?

In my work, I develop a **sound foundational understanding of the necessary and sufficient conditions for the reliability of ML models**. I leverage this understanding to **build reliable yet performant practical ML systems across diverse data modalities**. I revisit standard assumptions for learning by considering an *adversary*, which mathematically captures undesirable conditions such as poisoned training data, perturbed test data, out-of-distribution inputs, and label noise, to name a few. I take a comprehensive approach that starts by using an analytical lens to modeling vulnerabilities in ML systems in practice. The lack of reliability I find leads me to develop an understanding the conditions under which ML models can be reliable, and what properties these reliable ML models possess, focused on test-time robustness. Finally, I use these insights to develop techniques for embedding reliability in ML systems, going beyond robustness to address naturally-arising reliability concerns.

My key contributions to this research agenda are organized around four thrusts:

1. **Modeling realistic adversaries for practical ML models:** My work was the first to show how powerful test-time attacks can be implemented in practice, using numerical approximations of gradients [7] to attack a deployed ML system. I introduced a new training-time attack called model poisoning [2] on federated learning via stealthy modifications of model parameters. I have demonstrated physically-realizable attacks at both training [26, 25] and test time [24, 21], and that unreliable behavior can be induced even in out-of-distribution detectors [20, 8].
2. **Finding data-driven fundamental limits on the robustness of ML models:** I have shown that achievable bounds on the loss incurred by *any* classifier in the presence of a test-time adversary and for a given distribution can be found [3, 4, 12]. I utilized the theory of optimal transport to provide information-theoretic limits on robustness [3]. For the special case of empirical distributions, I introduced the concept of a conflict graph, enabling the use of graph algorithms to efficiently find lower bounds on the cross-entropy loss [4] and in the multi-class setting [12].
3. **Model-focused characterization of reliable behavior:** Using tools that analyze the internal representations of neural networks, I provided a cross-adversary analysis of how reliability manifests via different methods [10]. These internal representations are also amenable to conflict graph-based analysis, providing bounds on the robustness of different feature extractors [6]. Sample complexity in the presence of an adversary is important to understand training dynamics and I showed it is controlled by a generalization of the VC-dimension known as the Adversarial VC-dimension [11].
4. **Building reliable ML systems under different adverse conditions:** I have found forcing adversaries to make tradeoffs between incompatible goals is often the key to enhancing reliability. I have improved robustness in the face of test-time attacks for computer vision [5, 27, 13], tampered training data in vision and language tasks [18, 23]. My recent work has aimed at improving the reliability of ML systems deployed to solve problems of societal interest, particularly in networking. I have tackled issues of label noise in network probes for censorship detection [9], data drift in temporal cellular network data [16] and data scarcity for network traffic classification [14].

Research Impact: I strive to create impact on three fronts: academic, practical and interpersonal. My research has over 7000 citations and has been recognized through the Siemens FutureMakers Fellowship, a finalist position for the Bell Labs Prize, a $150k grant from C3.ai, a Spotlight paper at the NeurIPS conference, among others. The practical impact of my work stems from my collaborations with industry researchers on two monographs on reliable federated learning [15, 1], as well as Verizon's deployment of a solution for reliable resource forecasting in the presence of drift [16]. I believe strongly that mentorship is the key to impactful research. I have mentored 10 graduate and 3 undergraduate students as a researcher at the University of Chicago, guiding several to their first research publication.

Research Thrust 1: Identifying Unreliability in ML Systems

In my work, I have created powerful, theory-guided attacks that work in practical settings. I show how tools from numerical and regularized optimization can be used to find vulnerabilities in ML models.

Query-based black-box attacks: ML models are susceptible to adversarial examples, which are test inputs that have been strategically perturbed to induce incorrect inferences. Typically, these are generated by performing gradient descent with respect to the input on a loss function defined such that minimizing the loss corresponds to changing the model's prediction, while constraining the added perturbation to lie within a neighborhood of the original input.

In practice however, Machine Learning as a Servive (MLaaS) providers only allow input-output access to their models through an API. I found that this level of access, termed *black-box* since there is no access to the model's internals, is *sufficient to generate effective and stealthy adversarial examples against state-of-the-art models* [7]. I estimated gradients using the method of finite differences and reduced the number of queries drastically through dimensionality reduction.

Physically realizable attacks at training and test time: ML systems become even more unreliable when their training phase is compromised with malicious data. Backdoor attacks modify input training data with artifacts (called triggers) designed to make the model learn spurious correlations that are then exploited at test time. These attacks were limited to digital modifications until my collaborators and I showed that physical objects in images could be used as backdoor triggers [26, 25]. My research has also showed that adversarial examples can be effective even when encountered in the physical world, and passed through a sensor such as a camera by modifying the optimization process with an expectation over transformations that simulate possible artifacts [24], although different portions of a physical object have varying impact on the output [21]. I showed that even models designed to operate in an open-world setting are vulnerable, with the out-of-distribution detectors used to detect irrelevant inputs being brittle [20, 8].

Vulnerabilities in distributed learning: The lack of reliability is even more pernicious in decentralized modes of training such as federated learning, which is widely deployed in practice. I was the first to show *model poisoning* [2] is possible in federated learning, with a small number of agents participating in the training process being compromised and returning model updates that are modified to ensure that certain input data is misclassified. Regularized optimization was used to embed the malicious updates within benign ones, making compromised agents hard to detect.

Impact Highlights: My paper on query-based black-box attacks was the first to demonstrate the impact of adversarial examples on deployed, real-world systems with an attack on Clarifai's content moderation model. Model poisoning has emerged as a novel attack vector that has been extensively studied in follow-up work, with over 900 citations.

Research Thrust 2: Data-focused Fundamental Limits on Robustness

My work has shown that is possible to derive model-agnostic fundamental limits on robustness in the presence of an adversary. These bounds step away from an attack-defense arms race, and characterize how well the best model performs in the presence of the strongest possible test-time adversary. In addition, these bounds are tight, *i.e.* there exist classifiers that can achieve these bounds, providing a benchmark for practitioners to train reliable models.

Optimal robust loss via optimal transport: When the set of allowed classifiers over a real vector space is all measurable functions, the best possible classifier will achieve the Bayes loss for general distributions, and zero loss for discrete distributions, *i.e.* it can separate data perfectly. However, this is no longer be true when data can be adversarially perturbed at test time. To reason about reliability in a model-agnostic fashion, we need to understand the geometry of the data distribution when perturbed. The theory of optimal transport allows for the determination of distribution-level costs using point-wise costs between points. I showed that by defining an appropriate adversarial point-wise cost function, the transport and classification problems can be related using Kantorovich duality, and the *optimal transport cost provides a lower bound on the optimal loss over all possible models in the presence of perturbations* [3]. This result holds for arbitrary distributions defined on Polish spaces and upper-hemicontinuous, closed, and non-empty perturbation neighborhoods. For cases of interest such as empirical and Gaussian distributions, the optimal transport cost can be efficiently computed, unlike the optimal loss itself, making the identification of this duality extremely effective. For Gaussian data, the paper also determines the sample complexity of the optimal classifier.

Lower bounds on robust loss for empirical distributions: For empirical distributions used in practice, I have derived improved methods to lower bound robust loss using a conflict graph that captures collisions between perturbed samples [4, 12]. Finding the optimal loss directly for the case of the cross-entropy loss [4] and multi-class classification [12] is very computationally expensive. Using these bounds, I characterized the performance of training methods designed to improve reliability against adversarial examples, and demonstrated a large gap to optimality.

Impact Highlights: Efficiently computable and achievable fundamental limits on robustness allow researchers to determine the gap to robustness for practical models. This line of work has been recognized with a finalist position for the Bell Labs Prize, a $150k grant from C3.ai and a Spotlight paper at the NeurIPS conference.

Research Thrust 3: Model-focused Reliability Analysis

The development of reliable models rests on understanding how specific model families behave under adverse conditions. In my work, I show how a variety of model properties such as convergence to optimality, discrimination power and learned features in deeper layers change when test-time adversaries must be accounted for.

Explaining robust feature learning in deep neural networks: Classifiers that attain some level of robustness against adversarial examples have been obtained using adversarial training proposed by Madry *et. al* [17] and its variants. During training, each sample is perturbed using a specified attack before its loss is computed. However, it is unclear how the features learned by these robust models differ from standard models and across attacks. I used Centered Kernel Alignment (CKA), to compare learned representations, both robust and not, across different models [10]. I found *stronger attackers consume more of the capacity of neural networks, leading to feature collapse* across layers as well as differential convergence speeds across layers. Surprisingly, the results also show different attackers can lead to similar robust features, indicating the way forward for models reliable against multiple attackers simultaneously. Each layer of models robust to test-time attacks can be treated as a new data representation, which enables use of the conflict graph idea for determining lower bounds on robustness to feature extractors [6]. This allows for the *precise determination of how robust the best classifier trained on top of a given feature extractor can be*. This is interesting for both transfer learning using robustly trained models as well as to determine architectural properties that detract from robustness.

Convergence to robust models: For a given model family, it is of interest to practitioners to determine how many samples it takes for a hypothesis to converge to the best possible robust hypothesis from that class. In the standard framework of statistical learning theory, the sample complexity is governed by the Vapnik-Chervonenkis (VC) dimension of the model family used. I derived an analogue of this quantity, termed the *Adversarial VC-dimension* [11], when the data is adversarially modified. For linear models, this work showed that learning to be robust to perturbations constrained by standard distance metrics in real-valued spaces does not take more samples than standard learning.

Impact Highlights: Finding the convergence rate to robust linear models in the presence of adversarially modified data closed an open problem posed by Schmidt *et. al* [19].

Research Thrust 4: Building Reliable ML Systems

I use the insights from my analysis of the undesirable behavior of current ML systems to propose new training and data pre-processing methods. I also address naturally-occurring unreliable behavior in deployed ML systems.

Robustness against test-time attacks: I have developed robust training mechanisms that seek to build resilience into the training phase using *data transformations* [5]. These are a generalization of standard regularization methods using Principal Component Analysis (PCA). The data is transformed such that the learned model exhibits stronger dependence on high-variance components, which contain more information. This was followed up by work focusing on protecting against a different type of attacker who focuses their perturbation in a patch, as opposed to the entire input. Using models with low receptive fields greatly reduced the impact of patch-based attacks [27]. In recent work, my collaborators and I showed how optimized out-of-distribution data can be used to create model versions that have minimal transferability of adversarial examples from one model to another, enabling longitudinal robustness [13].

Securing the training pipeline: Training-time attacks are harder to protect against as the amount of data that needs to be compromised to induce undesirable behavior is often small. I have overcome this challenge in two separate ways. To make federated learning more reliable against model poisoning, my collaborators and I built SparseFed [18], an at-scale system that sparsifies agent updates, leading to a *provable trade-off for malicious agents between detectability and effectiveness*. *Post facto* reliability against training-time attacks can be achieved by building a system for post-attack forensics that can reliably trace back which data points led to the undesirable behavior getting embedded [23].

Reliability against natural data variation in ML for networks: ML systems are often unreliable due to natural properties of the input data, particularly in domains such as networking where the underlying data generation process shifts often. My collaborators and I proposed LEAF [16] to adaptively select new training data guided by the error distribution over newly acquired data, *drastically improving prediction performance on a real-world cellular network dataset*. For tasks such as censorship detection, existing heuristics are unreliable and can have tremendous label noise. The reliability of DNS censorship detection using ML [9] can be greatly improved on large-scale, real-world datasets by fusing labels from disparate sources, demonstrating the need for task-aware techniques.

Impact Highlights: LEAF is being deployed by Verizon to provide better forecasts for resource allocation in cellular networks, particularly to deal with exogenous shocks.

Future Research Goals

Equitable and reliable access for all stakeholders is integral to cultivating broad-based societal trust in a technology as transformative as machine learning. In the future, my vision is to enable the *creation of reliable ML systems that everyone can trust*. My current research has already begun working toward this vision by focusing on building long-term reliability into existing systems, while accounting for vulnerabilities in new ML paradigms. I will expand the ambit of my research to address different stakeholders in the pipeline of building and deploying ML systems, particularly in critical areas like healthcare and cybersecurity. I believe it is critical to empower model owners, data brokers and end users to influence and navigate ML systems. I will pursue three key lines of research towards achieving my vision:

1. **Identifying fundamental vulnerabilities for new ML paradigms:** As generative ML models take centre-stage, diagnosing the conditions under which they are unreliable has become critical for their deployment, although this is often an afterthought. I will find vulnerabilities in these systems and develop the tools needed for the next generation of these models to be more reliable. In particular, I am interested in exploring what modifications to input data can disrupt generative models' capabilities in creating specified types of content. Further, as automated ML-based tools to regulate human behavior online become more prevalent, it is critical to audit their effectiveness. Leveraging insights from my analytical work to determine the failure modes of these systems, I aim to help enhance their ability to appropriately regulate online speech. I am curious to explore these questions and others as ML continues to evolve rapidly, all the while focusing on enhancing trust in ML systems.

2. **Reliability as a fundamental property:** As the next generation of ML systems is created and deployed, long-term robustness to vulnerabilities, new and old, must be treated as a key property from inception. Current methods for building robust ML systems are either too expensive, lack generalization to new threats, or both. To this end, I am continuing to explore the interplay between fundamental limits on robustness and reliable ML system design. Currently, I am investigating the use of soft labels from conflict graphs over empirical distributions to guide training. I am also interested in how the scope of adverse conditions studied can be expanded, providing new constraints under which systems must be made reliable. I am studying the fundamental limits of robustness in the presence of multiple test-time attackers. In tandem, I am undertaking empirical research guided by these limits to build systems that can respond to new vulnerabilities that arise over their lifetime, leading to continual reliability. I find the temporal aspect of ML system deployment to be a rich one for theoretical investigation, as reliability must be maintained over time. Reliable ML research can also increase everyday users' trust, by protecting from them pernicious uses of ML such as website fingerprinting attacks [22]. This insight drives my goal of building reliable, yet flexible, ML systems that end users can navigate according to their personal choices.

3. **Enabling continued growth for ML:** As ML systems consume ever-increasing amounts of data, we may be reaching a plateau in terms of the performance that can be obtained by just ingesting raw data. In addition, the data hungry nature of current state-of-the-art ML systems ensures that only the most powerful stakeholders tend to have access to and say in the development of new ML systems. I believe continued, democratized growth in ML is only possible if techniques are developed to obtain relevant new data in data scarce regimes such as healthcare and cybersecurity. I aim to develop methods for task-aware data acquisition that would enable stakeholders with limited data to explore new data sources while minimizing commitment, computational overhead and privacy risks. Making these methods resilient to manipulation is an important future step in deploying it widely. Synthetic data generation using generative models is also a promising method to overcome data scarcity. My collaborators and I found that creative modifications of diffusion models allows for high quality data generation even in domains with stricter data constraints such as networking [14]. Together, these data-centric approaches can drive the next generation of ML systems.

Funding and collaboration: I recognize that the key to achieving my future research goals is to maintain ongoing collaborations while building new ones. I intend to continue my theory-driven work with Prof. Daniel Cullina (Penn State) on determining fundamental limits of robustness, extending it to new adversaries and models. I have an abiding collaboration with my Ph.D. advisor Prof. Prateek Mittal (Princeton) on analyzing and building reliable systems, particularly for training and test-time robustness. I am also working with Prof. Ben Zhao (University of Chicago) on finding vulnerabilities for generative image models such as diffusion models and with Prof. Nick Feamster (University of Chicago) on overcoming data scarcity in networking as well as auditing the reliability of ML models for content moderation.

References

[1] **A. Bhagoji** and S. Chakraborty. Securing federated learning: Defending against poisoning and evasion attacks. In L. M. Nguyen, T. N. Hoang, and P.-Y. Chen, editors, *Federated Learning: Theory and Practice*. Elsevier, 2024.

[2] **A. Bhagoji**, S. Chakraborty, P. Mittal, and S. Calo. Analyzing federated learning through an adversarial lens. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

[3] **A. Bhagoji**, D. Cullina, and P. Mittal. Lower bounds on adversarial robustness from optimal transport. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[4] **A. Bhagoji**, D. Cullina, V. Sehwag, and P. Mittal. Lower bounds on cross-entropy loss in the presence of test-time adversaries. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

[5] **A. Bhagoji**, D. Cullina, C. Sitawarin, and P. Mittal. Enhancing robustness of machine learning systems via data transformations. In *52nd Annual Conference on Information Sciences and Systems (CISS)*, 2018.

[6] **A. Bhagoji**, D. Cullina, and B. Y. Zhao. A theoretical perspective on the robustness of feature extractors. In *The Second Workshop on New Frontiers in Adversarial Machine Learning*, 2023.

[7] **A. Bhagoji**, W. He, B. Li, and D. Song. Practical black-box attacks on deep neural networks using efficient query mechanisms. In *European Conference on Computer Vision (ECCV)*, 2018.

[8] **A. Bhagoji** and P. Shirani. Adversarial attacks on anomaly detection. In D. Phung, G. I. Webb, and C. Sammut, editors, *Encyclopedia of Machine Learning and Data Science*. Springer US, 2020.

[9] J. Brown, X. Jiang, V. Tran, **A. Bhagoji**, N. P. Hoang, N. Feamster, P. Mittal, and V. Yegneswaran. Augmenting rule-based dns censorship detection at scale with machine learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2023.

[10] C. Cianfarani, **A. Bhagoji**, V. Sehwag, B. Zhao, H. Zheng, and P. Mittal. Understanding robust learning through the lens of representation similarities. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[11] D. Cullina, **A. Bhagoji**, and P. Mittal. Pac-learning in the presence of evasion adversaries. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

[12] S. Dai, W. Ding, **A. Bhagoji**, D. Cullina, B. Y. Zhao, H. Zheng, and P. Mittal. Characterizing the optimal 0-1 loss for multi-class classification with a test-time attacker. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[13] W. Ding, **A. Bhagoji**, B. Y. Zhao, and H. Zheng. Towards scalable and robust model versioning. In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2024.

[14] X. Jiang, S. Liu, A. Gember-Jacobson, **A. Bhagoji**, P. Schmitt, F. Bronzino, and N. Feamster. Netdiffusion: Network data augmentation through protocol-constrained traffic generation. *Accepted with Shepherding in Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 2024.

[15] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, **A. Bhagoji**, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 2021.

[16] S. Liu, F. Bronzino, P. Schmitt, **A. Bhagoji**, N. Feamster, H. G. Crespo, T. Coyle, and B. Ward. Leaf: Navigating concept drift in cellular networks. *Proceedings of the ACM on Networking*, 2023.

[17] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2017.

[18] A. Panda, S. Mahloujifar, **A. Bhagoji**, S. Chakraborty, and P. Mittal. Sparsefed: Mitigating model poisoning attacks in federated learning with sparsification. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

[19] L. Schmidt, S. Santurkar, D. Tsipras, K. Talwar, and A. Madry. Adversarially robust generalization requires more data. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

[20] V. Sehwag, **A. Bhagoji**, L. Song, C. Sitawarin, D. Cullina, M. Chiang, and P. Mittal. Analyzing the robustness of open-world machine learning. In *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security (AISec)*, 2019.

[21] V. Sehwag, C. Sitawarin, **A. Bhagoji**, A. Mosenia, M. Chiang, and P. Mittal. Not all pixels are born equal: An analysis of evasion attacks under locality constraints. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2018.

[22] S. Shan, **A. Bhagoji**, H. Zheng, and B. Y. Zhao. A real-time defense against website fingerprinting attacks. In *Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security (AISec)*, 2021.

[23] S. Shan, **A. Bhagoji**, H. Zheng, and B. Y. Zhao. Poison forensics: Traceback of data poisoning attacks in neural networks. In *31st USENIX Security Symposium (USENIX Security)*, 2022.

[24] C. Sitawarin, **A. Bhagoji**, A. Mosenia, P. Mittal, and M. Chiang. Rogue signs: Deceiving traffic sign recognition with malicious ads and logos. In *Deep Learning and Security Workshop (co-located with IEEE Security and Privacy)*, 2018.

[25] E. Wenger, R. Bhattacharjee, **A. Bhagoji**, J. Passananti, E. Andere, H. Zheng, and B. Zhao. Finding naturally occurring physical backdoors in image datasets. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[26] E. Wenger, J. Passananti, **A. Bhagoji**, Y. Yao, H. Zheng, and B. Y. Zhao. Backdoor attacks against deep learning systems in the physical world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[27] C. Xiang, **A. Bhagoji**, V. Sehwag, and P. Mittal. Patchguard: A provably robust defense against adversarial patches via small receptive fields and masking. In *30th USENIX Security Symposium (USENIX Security)*, 2021.