Research Statement: Towards *Verifiable* and Trustworthy Artificial Intelligence

Modern artificial intelligence (AI) has been broadly applied with deep neural networks as one of the most successful approaches. But the use of neural networks (NNs) in AI is often challenged in mission-critical settings such as cybersecurity, autonomous systems, and medical procedures because AI models are often "black boxes" and untrustworthy: it is hard to guarantee that they behave safely and predictably. Researchers have often identified surprisingly incorrect behaviors of AI models, and their brittleness may lead to unexpected catastrophic failures in real-world scenarios. This is a significant roadblock to the application of AI in scenarios where social responsibility is vital and risks are high.

My research aims to tackle these challenges by building trustworthy AI with *provably* safe and reliable behaviors in mission-critical and high-security tasks. My unique approach is to **develop formal verification methods for AI** and *rigorously* verify their trustworthiness. My work is among the **first to efficiently verify the trustworthiness of large neural networks** according to formal specifications, which was impossible with traditional verification methods such as satisfiability modulo theories due to the large model size. Theoretically, my bound-propagation-based verification framework can *formally guarantee* that a NN satisfies robustness, safety, or other specifications under malicious or adversarial input perturbations. Practically, to address the high verification cost, I proposed a series of specialized algorithms with high efficiency and scalability and developed a GPU-accelerated open-source verification toolbox. Finally, I showed how to build trustworthy AI models to achieve provable guarantees via theoretically principled training.

My verification algorithms for AI have become indispensable in this field, with key papers on neural network verification **cited over 1,000** times combined. My verification toolbox, α,β-CROWN, **won two consecutive years** of the International Verification of Neural Network Competition (VNN-COMP 2021, 2022), outperforming tools developed by Stanford, Oxford, UIUC, and other prestigious universities. Notably, in VNN-COMP 2022, the second-place tool also took a similar bound-propagation approach inspired by my strong results in 2021. My algorithms enable the verification of large neural networks with up to **millions of parameters**, and their scalability opens up the opportunity to apply formally verified AI in real-world mission-critical applications. My toolbox has been **used by large industrial players** such as Airbus and Collins Aerospace, and my representative verification algorithm (CROWN) has become a fundamental tool in solving challenging problems in other domains [19] and has been **taught in an MIT course** (16.332).

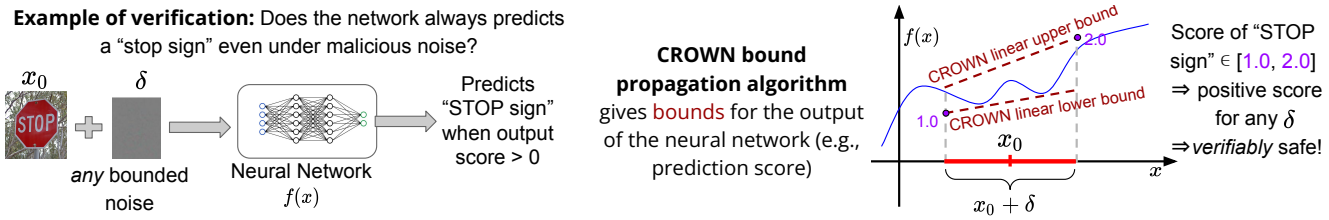
My future research starting in 2023 will be partly supported by the **Schmidt Futures AI2050 Early Career Fellowship** with a **$300,000** grant. Going forward, I aim to bring formal verification into a broad range of applications of AI, such as cybersecurity, robotics, autonomous systems, aeronautics, healthcare, and finance. I will also focus on further improving the scalability and strength of verifiers to enable verified AI in previously infeasible settings. Moreover, I will borrow insights from my successful NN verifiers to accelerate a broader class of discrete optimization problems.

1 Completed Research

My research is centered around *verifiable* and trustworthy AI, with three main themes detailed in this section:

- I proposed the **novel bound-propagation-based verification framework for neural networks**, which is scalable to large NNs and achieves up to three orders of magnitude speedup compared to generic solvers. My algorithms tackle verification as an optimization problem, exploit its structure for efficiency, and are amenable to GPU acceleration.
- To complement verification, I revealed **safety and security** issues in AI via **falsification methods** and attacks.
- I developed principled **training approaches to building robust AI**, including robust reinforcement learning (RL) under observational noises and robust classifiers using NNs or tree ensembles with *verifiable* robustness guarantees.

① **The bound propagation framework for formal verification of neural networks (NNs).** Formal verification of NNs aims to rigorously prove specifications involving NN output behavior (e.g., classification is correct) under input constraints (e.g., inputs with bounded malicious noises). In its canonical form, one must produce *sound lower and upper bounds* of NN outputs given arbitrary model inputs within constraints. Although traditional computer-aided verification techniques such as satisfiability modulo theories (SMT) or mixed integer programming (MIP) can be applied, these generic approaches can hardly scale to realistic models; verifying a NN with hundreds of neurons may take days.



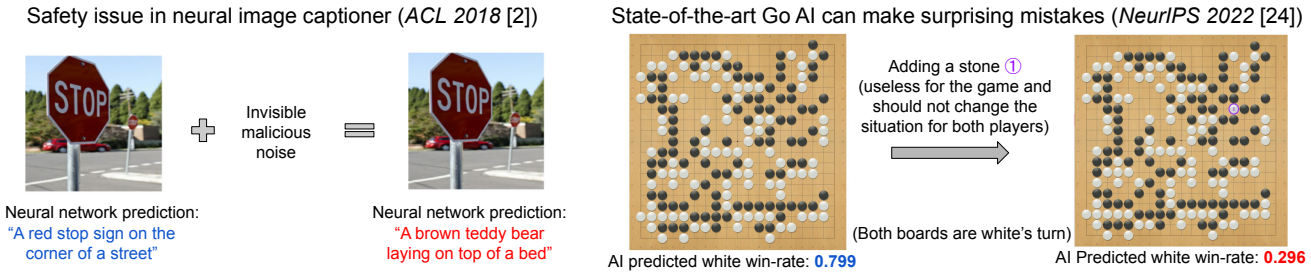
My first contribution is a **bound-propagation-based verification algorithm**, CROWN [5, 4], that can efficiently compute the lower and upper bounds of NNs given input perturbations. It is based on the key observation that non-linear activation functions can be replaced by their linear bounds, and these bounds can be carefully propagated through each layer to obtain a set of sound linear inequalities bounding the output of a non-convex NN function. Unlike generic SMT-based or MIP-based verification approaches, the bound propagation process *exploits the highly structured verification problem* with NNs in its backbone and can efficiently run on GPUs [15]. Theoretically, I showed that CROWN is equivalent to solving linear programming (LP) relaxed from MIP for ReLU networks [9], but crucially, CROWN exploits the NN structure by propagating linear bounds, can be accelerated on GPUs and scales much better than an LP solver.

My second contribution is to **generalize bound-propagation-based verification** to general computation functions and **widen their applicability**. Early verification algorithms were limited to feed-forward networks, and to verify a new architecture such as Transformer, I had to manually derive and implement verification bounds [13]. This burden greatly restricted the application of verification. To address this challenge, I extended CROWN to *general computation graphs*, including general NN architectures and computations [15]. My "auto_LiRPA" library is the first automatic toolbox giving tight lower and upper bounds under input perturbations for a computation function defined in PyTorch. The algorithmic efficiency of bound propagation allows auto_LiRPA to verify large models such as a 20-layer ResNet, while many verifiers at that time could be applied only to a few fully-connected layers. In addition, auto_LiRPA enables bound propagation for more problems by treating a complex function (such as a Jacobian) as a computation graph, resulting in new approaches to computing tight local Lipschitz constants and verifying monotonicity of NNs [25, 11].

My third contribution is to **strengthen the bound propagation formulation** via *branch-and-bound* and *cutting plane methods* and enable novel optimization methods to tighten verification. I identified the limitations of CROWN and many other verifiers (the "convex relaxation barrier" [9]). To break this barrier, β-CROWN [21] utilizes branch-and-bound (B&B) to strategically divide the verification problem into many subproblems [22] and conducts strengthened bound propagation on each subproblem. Unlike MIP/SMT solvers, B&B in β-CROWN utilizes bound propagation on GPUs and quickly enumerates millions of subproblems to verify challenging problems. Moreover, GCP-CROWN [26] extends bound propagation to include any general cutting plane constraints to further tighten the bound. β-CROWN and GCP-CROWN introduce novel optimizable parameters during bound propagation, allowing the use of fast gradient descent to tighten bounds while maintaining soundness. As a result, GCP-CROWN, the strongest bound-propagation-based algorithm, can completely solve all instances in `oval20` (a representative benchmark) with an average time of **3.5s**; in contrast, a generic MIP-based approach can solve only half of all instances with an average time of over **2,000s** [26].

With all these innovations above, to maximize the societal impact, I built a practical tool for practitioners to apply verification in domain-specific tasks. I led a multi-institutional team that developed an **open-source and award-winning neural network verifier**, α,β-CROWN, which won VNN-COMP 2021 and 2022, and can solve a variety of verification problems in computer vision, reinforcement learning, computer systems, and aerospace applications.

② **Security and falsification of artificial intelligence.** Besides verification, *falsification* aims to find counter-examples to disprove a certain formal specification. A practical NN verifier must either verify or falsify a given specification, so a strong falsification procedure is important. When the specification is robustness or safety, falsification is often referred to as an "adversarial attack", an active research topic in computer security. In this setting, the attacker or falsifier aims to find a slightly altered input that triggers an incorrect model behavior. I investigated the robustness of many different types of AI models, including image classification [3, 10], captioning [2], super-resolution [8], natural language classification [12], decision tree ensembles [16] and super-human AI agents playing Go [24]. These attacks are the first of their kind for these specific domains of AI, demonstrating the weaknesses of modern AI in many applications.



I also studied different formulations and threat models of adversarial attacks. ZOO-Attack [1] was the first to show that adversarial attacks of NNs can be conducted in the *query-based*, *black-box* setting, demonstrating the practicability of attacking real-world AI systems. ZOO-Attack is one of the earliest black-box attacks and is highly influential (**cited over 1,300 times**), which inspired a long line of papers on black-box attacks. I also proposed BaB-Attack [27], employing branch-and-bound for a systematic attack that can locate adversarial examples missed by all existing methods.

③ **Building robust and verifiable AI models via training.** Since the verification bounds computed by bound propagation are functions of network weights, one can use their gradients to update model weights and tighten the bounds. CROWN-IBP [17] is a specialized bound propagation algorithm for training, combining the tight CROWN bounds with cheap interval bound propagation, balancing training efficiency and bound effectiveness. It has become a standard baseline method for training verifiably robust NNs (or "certified adversarial defense"), and has influenced many works in this field (cited over 200 times). My recent work [20] further reduces the cost of training verifiable NNs, achieving formally verified robustness guarantees on relatively large models (wide ResNet, ResNeXt) and datasets (TinyImageNet). In addition, I also studied the effectiveness of using verification-aware model pruning to make NNs verifiable [23].

Beyond NNs and classification problems, I also studied how to build robust AI in other settings. For reinforcement learning, I developed the first theoretical framework, SA-MDP, to characterize agent behavior under adversaries on observations [18]. I proposed theoretically principled methods to train robust agents by regularizing a lower bound on reward [18] and alternatingly train an optimal adversary and the agent [16]. In addition, I studied the formal verification problem of tree ensembles [7], and proposed robust training of tree ensembles [6] with verifiable guarantees [14].

2 Future Directions

My future research aims to strengthen the capability and scalability of verifiers to enable trustworthy AI in previously infeasible scenarios, apply formally verified AI to more applications, and make broader impacts on related domains:

Verification beyond robustness. Although formal verification of NNs originated from robustness verification, the techniques I developed offered a principled way for verifying any computation graph, and verification specifications can also be extended to more complex ones. My vision of AI verification is that it *must be applicable to a large number of applications* (e.g., cyber-security, robotics, aeronautics, manufacturing, and healthcare) to make a bigger societal impact. I will collaborate with domain experts and precisely formulate the specifications arising from different domains, including different notions of trustworthiness: safety, stability, reliability, fairness, consistency, privacy, and others. Then, I will create novel algorithms (based on or beyond the bound propagation approach) that can handle these new formal specifications. For example, using AI to control a surgical robot (and many other cyber-physical and autonomous systems) requires safety and stability guarantees. Control theory requires the design of a barrier function satisfying a set of mathematical conditions describing safety scenarios, but they are challenging to solve for non-linear systems with complex NN-based controllers. My algorithms can help to find and prove these conditions by treating them as *general computation graphs* and verify the safety or stability using bound-propagation techniques. The good scalability of my approaches can enable provably stable and safe NN-based controllers in large and complex systems.

Strong and scalable verification via high-order bound propagation and verification-friendly system design. Making the verifier stronger and more scalable is crucial for applying verification-based techniques in demanding real-world applications. The tightness and scalability of verifiers can be improved from different perspectives. To approach a

new generation of verifiers, I will study a *new bound propagation framework* using high-order relaxations (such as polynomials); they may bypass the limitations of existing formulations of linear bound propagation and produce much tighter bounds, while still keeping the benefit of *fast bound propagation properties on general computation graphs* with a careful design, allowing stronger verification without sacrificing efficiency. To scale verification to large AI models, I will *exploit the flexibility that AI models can be trained or designed to achieve better verifiability*, and propose new training techniques and novel system architectures that are verification-friendly without sacrificing performance. Lastly, I aim to further develop the award-wining α,β-CROWN verification toolbox with additions of new algorithms developed in previous steps, and make this tool a strong and universal one applicable to different domain problems.

Accelerating optimization for a broader class of discrete optimization problems. The success of my NN verification techniques relies on the fast and GPU-accelerated bound-propagation solver that exploits the structure of the underlying optimization problem, combined with a highly parallel branch-and-bound process. Inspired by its success, I will extend the paradigm of bound propagation to other discrete optimization problems with good structures, possibly optimization problems on graphs. My approach is to develop a fast and relaxed solver for the underlying optimization problem that can be efficiently accelerated on GPUs in the same spirit as bound-propagation methods. The relaxed solver should be amenable to highly parallel branch-and-bound to strengthen its power to obtain tighter results, replacing existing MIP or SMT procedures. The goal of this direction aims to achieve a significant speedup on hard discrete optimization problems in engineering and science, similar to the success I have achieved in neural network verification.

To conclude, I have built a novel framework for the formal verification of neural networks and also thoroughly studied the safety and robustness of many AI models. I aim to further widen the impact of AI verification and build trustworthy and verifiable AI models in different domains. I envision a future where trustworthy AI models are widely deployed, enabling learning-based building blocks in mission-critical systems to enhance their performance.

References (* indicates **co-first** authors)

[1] Pin-Yu Chen*, **Huan Zhang***, Yash Sharma, Jinfeng Yi, Cho-Jui Hsieh. ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. *10th ACM Workshop on Artificial Intelligence and Security* (2017).

[2] Hongge Chen*, **Huan Zhang***, Pin-Yu Chen, Jinfeng Yi, Cho-Jui Hsieh. Attacking visual language grounding with adversarial examples: A case study on neural image captioning. *56th Annual Meeting of the Association for Computational Linguistics (ACL)* (2018).

[3] Dong Su*, **Huan Zhang***, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, Yupeng Gao. Is robustness the cost of accuracy?–a comprehensive study on the robustness of 18 deep image classification models. *Proceedings of the European Conference on Computer Vision (ECCV)* (2018).

[4] Tsui-Wei Weng*, **Huan Zhang***, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Luca Daniel, Duane Boning, Inderjit Dhillon. Towards fast computation of certified robustness for relu networks. *International Conference on Machine Learning (ICML)* (2018).

[5] **Huan Zhang***, Tsui-Wei Weng*, Pin-Yu Chen, Cho-Jui Hsieh, Luca Daniel. Efficient neural network robustness certification with general activation functions. *Advances in Neural Information Processing Systems (NeurIPS)* (2018).

[6] Hongge Chen, **Huan Zhang**, Duane Boning, Cho-Jui Hsieh. Robust decision trees against adversarial examples. *International Conference on Machine Learning (ICML)* (2019).

[7] Hongge Chen*, **Huan Zhang***, Si Si, Yang Li, Duane Boning, Cho-Jui Hsieh. Robustness verification of tree-based models. *Advances in Neural Information Processing Systems (NeurIPS)* (2019).

[8] Jun-Ho Choi, **Huan Zhang**, Jun-Hyuk Kim, Cho-Jui Hsieh, Jong-Seok Lee. Evaluating robustness of deep image super-resolution against adversarial attacks. *International Conference on Computer Vision (ICCV)* (2019).

[9] Hadi Salman, Greg Yang, **Huan Zhang**, Cho-Jui Hsieh, Pengchuan Zhang. A convex relaxation barrier to tight robustness verification of neural networks. *Advances in Neural Information Processing Systems (NeurIPS)* (2019).

[10] **Huan Zhang***, Hongge Chen*, Zhao Song, Duane Boning, Inderjit S Dhillon, Cho-Jui Hsieh. The limitations of adversarial training and the blind-spot attack. *International Conference on Learning Representations (ICLR)* (2019).

[11] **Huan Zhang**, Pengchuan Zhang, Cho-Jui Hsieh. RecurJac: An efficient recursive algorithm for bounding Jacobian matrix of neural networks and its applications. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)* (2019).

[12] Minhao Cheng, Jinfeng Yi, Pin-Yu Chen, **Huan Zhang**, Cho-Jui Hsieh. Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)* (2020).

[13] Zhouxing Shi, **Huan Zhang**, Kai-Wei Chang, Minlie Huang, Cho-Jui Hsieh. Robustness verification for transformers. *International Conference on Learning Representations (ICLR)* (2020).

[14] Yihan Wang, **Huan Zhang**, Hongge Chen, Duane Boning, Cho-Jui Hsieh. On Lp-norm robustness of ensemble stumps and trees. *International Conference on Machine Learning (ICML)* (2020).

[15] Kaidi Xu*, Zhouxing Shi*, **Huan Zhang***, Yihan Wang, Kai-Wei Chang, Minlie Huang, Bhavya Kailkhura, Xue Lin, Cho-Jui Hsieh. Automatic perturbation analysis for scalable certified robustness and beyond. *Advances in Neural Information Processing Systems (NeurIPS)* (2020).

[16] Chong Zhang, **Huan Zhang**, Cho-Jui Hsieh. An efficient adversarial attack for tree ensembles. *Advances in Neural Information Processing Systems (NeurIPS)* (2020).

[17] **Huan Zhang**, Hongge Chen, Chaowei Xiao, Bo Li, Duane Boning, Cho-Jui Hsieh. Towards stable and efficient training of verifiably robust neural networks. *International Conference on Learning Representations (ICLR)* (2020).

[18] **Huan Zhang***, Hongge Chen*, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, Cho-Jui Hsieh. Robust deep reinforcement learning against adversarial perturbations on state observations. *Advances in Neural Information Processing Systems (NeurIPS)* (2020).

[19] Michael Everett. Tutorial on safety verification and stability analysis of neural network-driven systems. *IEEE Conference on Decision and Control (CDC)* (2021).

[20] Zhouxing Shi*, Yihan Wang*, **Huan Zhang**, Jinfeng Yi, Cho-Jui Hsieh. Fast certified robust training via better initialization and shorter warmup. *Advances in Neural Information Processing Systems (NeurIPS)* (2021).

[21] Shiqi Wang*, **Huan Zhang***, Kaidi Xu*, Xue Lin, Suman Jana, Cho-Jui Hsieh, J. Zico Kolter. β-CROWN: Efficient bound propagation with per-neuron split constraints for neural network robustness verification. *Advances in Neural Information Processing Systems (NeurIPS)* (2021).

[22] Kaidi Xu*, **Huan Zhang***, Shiqi Wang, Yihan Wang, Suman Jana, Xue Lin, Cho-Jui Hsieh. Fast and complete: Enabling complete neural network verification with rapid and massively parallel incomplete verifiers. *International Conference on Learning Representations (ICLR)* (2021).

[23] Tianlong Chen*, **Huan Zhang***, Zhenyu Zhang, Shiyu Chang, Sijia Liu, Pin-Yu Chen, Zhangyang Wang. Linearity grafting: Relaxed neuron pruning helps certifiable robustness. *International Conference on Machine Learning (ICML)* (2022).

[24] Li-Cheng Lan, **Huan Zhang**, Ti-Rong Wu, Meng-Yu Tsai, I Wu, Cho-Jui Hsieh. Are AlphaZero-like agents robust to adversarial perturbations? *Advances in Neural Information Processing Systems (NeurIPS)* (2022).

[25] Zhouxing Shi, Yihan Wang, **Huan Zhang**, J. Zico Kolter, Cho-Jui Hsieh. Efficiently computing local Lipschitz constants of neural networks via bound propagation. *Advances in Neural Information Processing Systems (NeurIPS)* (2022).

[26] **Huan Zhang***, Shiqi Wang*, Kaidi Xu*, Linyi Li, Bo Li, Suman Jana, Cho-Jui Hsieh, J. Zico Kolter. General cutting planes for bound-propagation-based neural network verification. *Advances in Neural Information Processing Systems (NeurIPS)* (2022).

[27] **Huan Zhang***, Shiqi Wang*, Kaidi Xu, Yihan Wang, Suman Jana, Cho-Jui Hsieh, J. Zico Kolter. A branch and bound framework for stronger adversarial attacks of relu networks. *International Conference on Machine Learning (ICML)* (2022).