Shaping the Future of AI: Ensuring Robustness, Efficiency, and Fairness in Evolving Models

As we forge ahead into an era of rapidly evolving Artificial Intelligence and Machine Learning, Large Language Models (LLMs), Vision Language Models (VLMs), and generative AI are becoming increasingly intertwined with our lives. These powerful tools hold the potential to revolutionize countless domains - from healthcare to transportation, education to entertainment, our workspaces to our homes. But this immense potential does not come without its perils. We have witnessed instances where AI/ML models have fallen short of our expectations due to lack of robustness, efficiency, and fairness. For instance, the AI chatbot 'Tay' by Microsoft began to parrot offensive and inappropriate content, becoming a striking example of AI susceptibility to spurious features. Similarly, self-driving cars have shown vulnerability to adversarial perturbations - simple stickers strategically placed on stop signs have deceived these AI models into misclassifying them. Moreover, many AI models have faltered when faced with distribution shifts, failing to generalize their learning from training to real-world conditions, as evidenced by AI's frequently documented struggles with recognizing faces from underrepresented groups. These models' efficiency is another critical concern in the age of proliferating AI applications. With computational resources and data privacy being significant constraints, we need models that are lean and data-efficient. The recent Transformer models, despite their impressive capabilities, are notorious for their demand for computational resources and extensive training data, bringing us to a pressing need for efficient model design, data utilization, and learning processes. Additionally, as AI models continue to influence decision-making in crucial areas like healthcare, recruitment, and law enforcement, fairness has emerged as a non-negotiable requirement. Long-term fairness is especially challenging to achieve, as these AI systems often encounter evolving data distributions over time, which may lead to a drift from their fairness criteria.

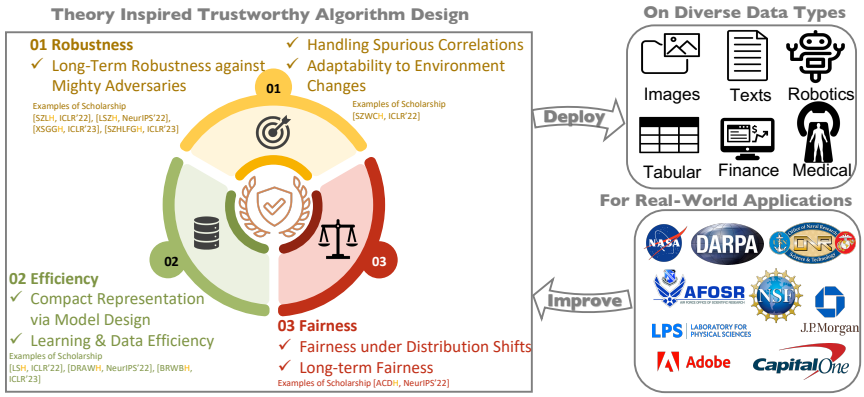

Figure 1: My research cycle: design theory inspired trustworthy AI/ML algorithms, deploy on diverse data types, and get feedback from real-world application to improve/inspire theory and algorithms.

In the face of these challenges, my research stands at the forefront, focusing on **robustness**, **efficiency**, and **fairness** in AI/ML models, vital in fostering an era of Trustworthy AI that society can rely on. My research fortifies models against spurious features, adversarial perturbations, and distribution shifts, enhances model, data, and learning efficiency, and ensures long-term fairness under distribution shifts. My pioneering works on tensor methods, to improve efficiency and robustness of learning, have not only advanced our understanding of tensor methods but also opened up new possibilities for their application in AI/ML. My research has made fundamental contributions to non-convex optimization, which lays theoretical foundations for optimization in deep learning and sparks a surge of followup works. My research has culminated in 111 papers (11 in the year of 2021, 27 in 2022, 30 in 2023, 13 in 2024 as of March) published in highly regarded venues (40 in NeurIPS, 27 in ICLR and 18 in ICML). The extensive citations these publications have received highlight the resonance of these contributions within the AI/ML community and beyond, recognized by 2 best/outstanding paper awards, winner of MIT TR35 Asia Pacific, finalist of AI researcher of the year 2022, 3 JP Morgan Faculty Research Awards, MLconf Industry Impact Award, etc. With academic and industrial collaborators, my research has been used for cataloguing brain cell types, learning human disease hierarchy, designing non-addictive pain killers, controlling power-grid for resiliency, defending against adversarial entities in financial markets, updating/finetuning industrial-scale model efficiently and etc. In the following sections, I will discuss my contributions and future plans in shaping robust, efficient, and fair AI/ML models in more detail.

Research Accomplishments

Focus 1: Robust and Adaptable Real-World Sequential Decision Making Systems.

Reinforcement learning creates intelligent agents capable of making automated sequential decisions. These agents have shown remarkable performance in stable and simulated environments such as video and board games. However, real-world deployment of these intelligent agents presents various challenges. One of these challenges is the prevalence of noise, false information, or adversarial inputs that can cause well-trained agents to fail. Another challenge is the ever-changing nature of the real world –intelligent agents struggle to adapt quickly to such changes as they require a large amount of data samples and computation to learn. My research aims to enhance the robustness and adaptability of decision-making agents in reinforcement learning, enabling them to withstand input variations and efficiently adjust to task or environment changes. Leveraging 'knowing the long-term vulnerability' and 'learning by analogy,' it enables robust learning against perturbations and knowledge transfer to new contexts. My study addresses two key facets of building reliable intelligent systems.

(1) Robustness: My research investigates the vulnerability of intelligent agents to spurious features/correlations, noisy or adversarial inputs and aims to develop agents that can perform reliably regardless of input perturbations.

(1a) Long-Term Robustness against Mighty Adversaries. I pioneered groundbreaking work, honored with a Best Paper Award, identifying and evaluating the vulnerabilities of ML-assisted, long-term decision-making systems against adversarial perturbations. Taking inspiration from Sun Tzu's "Art of War", we unveil a stark vulnerability of deep RL agents to long-term attacks [23] – an efficient strongest attack that outperforms previous adversarial attack methodologies. We then introduced a 'worst-case bellman operator [10],' a game-changer that trains robust RL agents, requiring no additional learning samples and ensuring state-of-the-art (SOTA) performance across a variety of attacks, SOTA Pareto frontier of robustness vs natural reward, and SOTA efficiency. Moreover, our recent approach [15] towards achieving robust neural networks through architectural design in the spectral domain, and our recent study, DyART [26] (a direct quantification of decision boundary movement to push decision boundary away from data points for adversarial robustness), ranking second on the RobustBench Leaderboard, all underscore our achievements in adversarial robustness. Pivoting from traditional ℓ_p-bounded attacks, we tackle multi-agent decision-making under arbitrary communication corruption, presenting a certifiably robust policy learning algorithm that enables explainable, certifiable decision-making [22]. Taking a more realistic angle, we looked at multi-agent RL scenarios where a victim agent is exploited by another agent controlled by an attacker, rather than the traditional direct manipulation of the victim agent. Our approach provides the first provably robust defenses with a convergence guarantee to the most robust victim policy, in sharp contrast to adversarial training in supervised learning which may only provide empirical defenses, earning an Outstanding Paper Award in recognition of this breakthrough work [11]. Finally, our work on tackling training-time perturbations or 'poisoning attacks' shines a light on the susceptibility of online, on-policy RL learners to these attacks. We proposed a practical algorithm that quantifies the robustness of RL agents against these threats [19], spurring further robustness studies.

(1b) Handling Spurious Correlations. We dive into the realm of representation learning, challenging the norm by not merely seeking statistical independence but striving for *causal disentanglement* of latent variables, allowing more controllable data generation, improved robustness, and better generalization. We champion a pioneering framework that models confounders leveraging domain expertise, to uncover the intricately woven causative factors - a stride beyond conventional approaches. This allows sufficient identification of the causally-disentangled factors while previous works can only discover necessary but not sufficient factors [13]. We confront the Achilles heel of existing models – robustness to *unforeseen data variations*, i.e., robustness to data variations without known characterization or training examples reflecting them, such as 3D-view or illumination change. Our strategy harnesses out-of-distribution data from divergent domains, utilizing an innovative equivariant domain translator. The result is a robust model, unfazed by even unexpected data changes, a leap forward in machine learning resilience [29]. We have broken the dependence on prior domain knowledge in data augmentation. Through an ingenious learning objective based on representation learning principles, we *truly autonomously learn augmentations* from scratch, different from prior autonomous methods that requires pre-specific pool of augmentations. Our method stands out, achieving groundbreaking results in complex areas such as Medical Images [27]. The techniques I have developed for enhancing the robustness of ML models have been adopted by other researchers and practitioners, significantly contributing to the creation of more reliable AI systems.

(2) Adaptability to Environment Changes: The research aims to develop agents that can adapt efficiently to changes in task specification or the underlying environment without the need for expensive and inefficient retraining.

Adapting by Knowledge Transfer. My work on transfer learning expedites autonomous decision-making systems in

dynamically changing environments with a theoretically guaranteed improvement of state-of-the-art effectiveness and efficiency. A key to achieving adaptability is knowledge transfer from the known to the unknown, while it is not easy to let agents automatically decide what to transfer and how to transfer, and avoid negatively affecting future performance. We improve the adaptability of agents by theoretically grounded knowledge transfer on various levels. On the first level, we focus on adapting to unseen states within a single environment, and we develop an algorithm to "learn by analogy" with guaranteed efficiency [18]. On the second level, we transfer knowledge from task to task, by proposing a provably efficient algorithm that utilizes the modular similarities across tasks [21]. Stepping towards another level, we let the agent adapt to drastically different observation spaces by transferring similarities of latent dynamics. We propose a theory-inspired transfer algorithm which, for the first time, achieves transfer learning from a vector-input environment to a pixel-input environment [24], providing a highly practical method for "sim2real" scenarios.

Adapting by Representation Pretraining. Pretraining is an increasingly prevalent direction to improve agents' adaptability. By mapping high-dimensional inputs into a lower-dimensional representation space while keeping the most informative features, one can greatly reduce the learning burden of downstream tasks. Our recent work, which designs a reward-free and world-model focused *foundation model for RL* [20], formulates the pretraining pipeline for RL tasks, and introduces a pretraining approach that leverages both perceptual and control-relevant information in the pretrained representation. The pretrained model can quickly adapt to multiple downstream tasks across various environmental structures.

Focus 2: Efficient, Adaptable and Generalizable Representation Learning.

Machine learning, especially reinforcement learning, has seen significant advancements in the past decade, fueled by high-performance computing and large datasets. However, large models trained on complex and rapidly growing data consume enormous computational resources. For example, a single training run of GPT-3 on 45 TB of text data takes a month on 1,024 A100 GPUs and costs $12M. Therefore, computationally-efficient algorithms for ML are urgently needed to continue making advancements in research, deployment, and accessible tooling to everyone.

(1) Compact Representation via Model Design. *Model design through Tensor Representation.* Our previous research [8, 9, 16, 14, 28] has demonstrated that theoretical insights on spectral methods and learning theory shed lights upon understanding of neural networks. Using tensor representation theory, we extend traditional neural networks to more general *tensorial neural networks* that generate interpretable and compact representations [14, 17]. We demonstrate tensorial neural networks' advantages over traditional neural networks by theoretically proving their improved generalization and expressive power with fewer parameters [9], invariances to group transformations [28] and robustness [15]. *A flexible design of Transformers.* We provide an interpretation of transformers with self-attention units based on which we introduce a compact design of self-attention units that achieves better than state-of-the-art performance using much fewer (e.g 1%) number of parameters in NLP, audio detection and computer vision tasks [12]. *Non-IID graph data.* We propose a principled and fundamentally different approach, VQ-GNN, a universal framework to scale up any convolution-based GNNs using Vector Quantization (VQ) without compromising the performance [4]. We propose the first sublinear time and memory framework for GNNs, through a sketch-based algorithm by training GNNs atop a few compact sketches of graph adjacency and node embeddings. Based on polynomial tensor-sketch (PTS) theory, our framework provides a novel protocol for sketching non-linear activations and graph convolution matrices in GNNs, as opposed to existing methods that sketch linear weights or gradients in neural networks. In addition, we develop a locality-sensitive hashing (LSH) technique that can be trained to improve the quality of sketches [6].

(2) Learning and Data Efficiency. *Nonconvex Optimization.* My research made fundamental contributions to non-convex optimization — the core tool used in deep learning, a prevalent AI/ML model. My collaborators and I are the first to prove that first-order gradient information, which is efficient to compute, guarantees convergence to (locally) optimal solutions for non-convex optimization [7]. This was surprising since all the earlier works needed second-order information, which is expensive to compute, for convergence. This pioneering work on non-convex optimization lays theoretical foundations for optimization in deep learning and sparks a surge of followup works. *Decentralized federated learning.* Centralized Federated learning provides an efficient collaborative learning paradigm, but has a single point of failure (the server) and its speed is blocked by slow workers. We propose decentralized federated learning with wait-free asynchronous SGD that no longer has a single point of failure and is no longer blocked by slow workers [2]. Theoretically, this asynchronous method matches the golden standard iteration convergence rate of parallel SGD, but (1) converges faster than synchronous parallel SGD as it doesn't require synchronization and (2) converges faster than asynchronous parallel SGD as it is no longer blocked by the slow worker. It achieves SOTA performance in extensive large-scale experiments. *Data condensation for energy-efficient architecture search.* Dataset condensation, aimed at reducing the computational load of training models on extensive datasets, often falters in terms of generalizability across varying hyperparameters/architectures. In our research, we pivot towards a con-

densation objective specifically crafted for hyperparameter search, aiming for synthetic validation datasets that mirror the performance rankings of models trained on original data under different hyperparameters. Introducing the innovative 'Hyperparameter-Calibrated Dataset Condensation' algorithm [5], we generate the synthetic validation dataset by matching hyperparameter gradients, achieved through implicit differentiation and efficient inverse Hessian approximation. Empirical results highlight the algorithm's proficiency in maintaining validation-performance rankings, while accelerating hyperparameter/architecture searches in image and graph tasks, thereby revolutionizing the dataset condensation field. My work on data efficiency has offered a pathway towards more sustainable AI development and utilization. Through these contributions, I am shaping the way the field thinks about and approaches the building of AI/ML models.

Focus 3: Ethical AI/ML.

Fairness under distribution shifts. As machine learning aids critical decisions in fields like hiring, loan approval, facial recognition, and criminal justice, there is a growing need for 'fair' models that do not discriminate based on protected attributes like race or gender. However, many existing fairness models may fail under distribution shifts. For example, a fair income predictor trained in one state may not be fair in another. We are the first to adapt a fair source model to a target domain, aiming for accuracy and fairness in both domains, despite these distribution shifts [1]. *Long-Term fairness.* Decisions made by machine learning models can have long-term impacts, and addressing these biases in sequential decision-making is crucial. Previous methods of summing static bias over time fail to consider state importance during transition, potentially causing a false sense of fairness. We introduce ELBERT, a new long-term fairness notion that considers state importance and preserves static fairness principles in sequential settings [25]. *Privacy.* Training models on private data risks user information leaks, particularly with generative models like chatbots. To prevent such leaks, we propose differentially private training methods, building on our differentially private spectral methods [3]. These methods learn from training data without memorizing specific user attributes, thus preventing data leakage. These advancements in fairness under distribution shifts, long-term fairness, and privacy protection underscore the broader impact of our research in promoting trustworthy AI. By creating AI models that are robust, fair, and protective of privacy, we're building a foundation for more ethical and equitable technology. These models have the potential to mitigate systemic biases, protect individual privacy, and ensure fairness across diverse applications, from hiring to healthcare. Our research lays the groundwork for a future where AI systems are a trusted and beneficial tool for society, underscoring our commitment to advancing the field towards responsible AI.

Fundraising Success, Teaching, Mentoring and Service

Throughout my academic career, I have demonstrated a strong ability to secure *funding*, amassing over $7M from 9 government grants and 6 private organizations, including 2 prestigious awards from the NSF (both as lead PI) and 7 from the DoD (6 of which as lead PI). I am passionate about *teaching* and have developed undergraduate and graduate machine learning courses, covering learning theory, latent variable models, and principled methods in deep learning and reinforcement learning. These courses have served 450 undergraduate students and 250 graduate students, for whom I have developed lecture content, assignments, and projects. I am deeply committed to fostering an inclusive learning environment for students of diverse backgrounds, a commitment that is particularly important given the lack of diversity in Computer Science and Machine Learning. *Advising/Mentoring.* I have served on 47 Dissertation Committees, chairing 2 of them. I have advised 20 undergraduates and 27 PhD students in their research, of which 7 undergraduates and 6 PhD students are female. In addition, I have mentored 5 female and 4 male undergraduates through UMD's NSF-funded REU program on learning fair representations of data without discrimination. In terms of *service and outreach*, I plan to build a partnership with STEM teachers from local high schools in the greater DC area to support early AI education and strengthen Computer Science pathways. I am also active in supporting women in Machine Learning, organizing the "Rising Stars in ML" event, and mentoring underrepresented students through UMD's "Center for Women in Computing" and the summer REU program. I organized the "Dagstuhl Seminar on Tensor Computations" in March of 2020 and 2021, bringing together researchers from various fields. Additionally, I organized a workshop on Matrix Factorization at the "Heidelberg Laureate Forum" and presented a tutorial on the method of moments and tensor decomposition. I co-organized an NSF-IEEE workshop on toward explainable, reliable, and sustainable machine learning in signal & data science in March 2023. I have also delivered a half-day tutorial on tensor methods at SIGMETRICS, the ACM Special Interest Group for the computer systems performance evaluation community. Moving forward, I plan to continue fostering collaborations within and beyond my institution and mentoring the next generation of researchers in developing trustworthy AI/ML models.

Future Research and Career Development: Aspirations for Continued Scholarly Impact

My future research plans remain committed to the pursuit of trustworthy AI. I plan to build on my existing research on robustness, efficiency, and fairness of AI/ML models. By delving deeper into these areas, I aim to uncover new insights and develop innovative solutions to the pressing challenges in the field, such as how to improve robustness/fairness and accuracy simultaneously without trading one for the other, how to achieve robustness more efficiently, and how to obtain fairness more efficiently. Beyond my current research focus, I am eager to explore new areas, such as generative AI, foundation models for robotics, within the realm of trustworthy AI. Potential directions include investigating the ethical implications of AI and developing mechanisms for better human-AI collaboration.

Direction 1: Detection of AI Generated Contents. The rise of generative AI models such as Language Models (LLMs), Diffusion Models, and Vision Language Models (VLMs) presents significant ethical concerns, notably their potential misuse. These powerful models could propagate disinformation, plagiarize content, generate false product reviews, or manipulate online discussions, leading to a distorted public perception of reality and potential harm to businesses and public discourse. To combat these issues, my future research will focus on developing mechanisms to differentiate between human- and machine-generated content. This essential work aims to mitigate the misuse of generative AI, ensuring its responsible and ethical application within society. Through this research, I aim to bolster the field of trustworthy AI, fostering a digital environment where AI supports human endeavours.

Direction 2: Copyright Protection. The proliferation of generative AI models, including Language Models (LLMs), Diffusion Models, and Vision Language Models (VLMs), has introduced significant challenges in the realm of intellectual property protection. The possibility of these models unwittingly forging or memorizing parts of the human-generated training data they learn from, particularly when this data consists of artistic or creative content, presents a substantial risk to copyright protection. In response to these challenges, I will work on: *(a) Preventing Data Forgery and Memorization:* Creating AI models that protect the originality of training data to guard against excessive memorization or replication. *(b) Credit Assignment Mechanism:* Establishing a fair revenue sharing system for data contributors, recognizing and compensating them appropriately. *(c) Safe Content Publicity:* Designing methods, such as data poisoning techniques, to help authors safely publicize their work without fear of misuse by AI models. My goal is to foster responsible and ethical AI use, where intellectual property rights are respected, contributing to a future where AI supports human creativity.

Direction 3: Trustworthy Foundation Models for Robot-Human Teaming. As AI progresses and begins to permeate more areas of our lives, particularly in the realm of robotics and control, there is a growing need for trustworthy foundation models that are not only robust to environmental shifts, noise, and perturbations, but can also effectively learn and adapt to human communication methods. I intend to construct foundational models for robotics that are pre-trained to understand world-models, prioritizing the capability of these models to quickly learn from humans. Recognizing that humans communicate with remarkable efficiency, often defining abstract terms to express complex concepts, I aim to capture this efficiency in the models I develop. Existing frameworks like imitation and reinforcement learning, which only allow for low-level symbol communication, fall short of replicating this aspect of human communication, leading to less efficient learning algorithms. In response, I propose to develop a novel framework that promotes progressively efficient learning and incorporates practical simulations of human teachers offering diverse, adaptive feedback. This innovative learning algorithm will foster the emergence of a human-like communication pattern, whereby the agent and the teacher streamline their interaction over time, exchanging increasingly abstract intentions. Through this endeavor, I aim to bridge the gap between humans and robots, paving the way for more effective, efficient, and meaningful robot-human teaming in a range of applications.

As a dedicated faculty member at a university, I remain committed to intertwining cutting-edge research with education, bringing the most recent technological content and crucial issues to the forefront of my students' learning. In my ongoing role as the Chapter Chair of the IEEE, by continuing to assume visible leadership roles in technology, I aim to inspire, engage, and help shape the future generation of AI and machine learning researchers, ensuring our field advances responsibly, inclusively, and innovatively.

References

[1] Bang An, Zora Che, Mucong Ding, and Furong Huang. Transferring fairness under distribution shifts via fair consistency regularization. In *Neural Information Processing System (NeurIPS)*, 2022.

[2] Marco Bornstein, Tahseen Rabbani, Evan Z Wang, Amrit Bedi, and Furong Huang. Swift: Rapid decentralized

federated learning via wait-free model communication. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[3] Chris Decarolis, Mukul Ram, Seyed Esmaeili, Yu-Xiang Wang, and Furong Huang. An end-to-end differentially private latent dirichlet allocation using a spectral algorithm. In *International Conference on Machine Learning (ICML)*, pages 2421–2431. PMLR, 2020.

[4] Mucong Ding, Kezhi Kong, Jingling Li, Chen Zhu, John Dickerson, Furong Huang, and Tom Goldstein. Vq-gnn: A universal framework to scale up graph neural networks using vector quantization. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021.

[5] Mucong Ding, Xiaoyu Liu, Tahseen Rabbani, and Furong Huang. Faster hyperparameter search on graphs via calibrated dataset condensation. In *New Frontiers in Graph Learning Workshop, NeurIPS*, 2022.

[6] Mucong Ding, Tahseen Rabbani, Bang An, Evan Wang, and Furong Huang. Sketch-gnn: Scalable graph neural networks with sublinear training complexity. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:2930–2943, 2022.

[7] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points — online stochastic gradient for tensor decomposition. In *Conference on Learning Theory (COLT)*, pages 797–842. PMLR, 2015.

[8] Furong Huang, Jordan Ash, John Langford, and Robert Schapire. Learning deep resnet blocks sequentially using boosting theory. In *Proceedings of the 35st International Conference on Machine Learning (ICML)*, 2018.

[9] Jingling Li, Yanchao Sun, Jiahao Su, Taiji Suzuki, and Furong Huang. Understanding generalization in deep learning via tensor methods. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 504–515. PMLR, 2020.

[10] Yongyuan Liang, Yanchao Sun, Ruijie Zheng, and Furong Huang. Efficient adversarial training without attacking: Worst-case-aware robust reinforcement learning. In *Neural Information Processing System (NeurIPS)*, 2022.

[11] Xiangyu Liu, Souradip Chakraborty, and Furong Huang. Controllable attack and improved adversarial training in multi-agent reinforcement learning. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS*, 2022.

[12] Xiaoyu Liu, Jiahao Su, and Furong Huang. Tuformer: Data-driven design of transformers for improved generalization or efficiency. In *The Tenth International Conference on Learning Representations (ICLR)*, 2022.

[13] Xiaoyu Liu, Jiaxin Yuan, Bang An, Yuancheng Xu, Yifan Yang, and Furong Huang. C-disentanglement: Discovering causally-independent generative factors under an inductive bias of confounder. In *The Sec- ond Workshop on Spurious Correlations, Invariance and Stability (SCIS) and Workshop on Structured Probabilistic Inference & Generative Modeling (SPIGM), ICML*, 2023.

[14] Alexander Reustle, Tahseen Rabbani, and Furong Huang. Fast gpu convolution for cp-decomposed tensorial neural networks. In *Proceedings of SAI Intelligent Systems Conference (IntelliSys)*, pages 468–487. Springer, 2020.

[15] Jiahao Su, Wonmin Byeon, and Furong Huang. Scaling-up diverse orthogonal convolutional networks with a paraunitary framework. In *International Conference on Machine Learning (ICML)*. PMLR, 2022.

[16] Jiahao Su, Milan Cvitkovic, and Furong Huang. Sampling-free learning of bayesian quantized neural networks. In *International Conference on Learning Representations (ICLR)*, 2020.

[17] Jiahao Su, Jingling Li, Xiaoyu Liu, Teresa Ranadive, Christopher Coley, Tai-Ching Tuan, and Furong Huang. Compact neural architecture designs by tensor representations. *Frontiers in Artificial Intelligence*, 5, 2022.

[18] Yanchao Sun and Furong Huang. Can agents learn by analogy? an inferable model for pac reinforcement learning. *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 2020.

[19] Yanchao Sun and Furong Huang. Vulnerability-aware poisoning mechanism for online rl with unknown dynamics. *The International Conference on Learning Representations (ICLR)*, 2020.

[20] Yanchao Sun, Shuang Ma, Ratnesh Madaan, Rogerio Bonatti, Furong Huang, and Ashish Kapoor. Smart: Self-supervised multi-task pretraining with control transformers. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[21] Yanchao Sun, Xiangyu Yin, and Furong Huang. Temple: Learning template of transitions for sample efficient multi-task rl. *AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[22] Yanchao Sun, Ruijie Zheng, Parisa Hassanzadeh, Yongyuan Liang?, Soheil Feizi, Sumitra Ganesh, and Furong Huang. Certifiably robust multi-agent reinforcement learning against adversarial communication. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[23] Yanchao Sun, Ruijie Zheng, Yongyuan Liang, and Furong Huang. Who is the strongest enemy? towards optimal and efficient evasion attacks in deep rl. In *The Tenth International Conference on Learning Representations (ICLR)*, 2022.

[24] Yanchao Sun, Ruijie Zheng, Xiyao Wang, Andrew E Cohen, and Furong Huang. Transfer rl across observation feature spaces via model-based regularization. In *The Tenth International Conference on Learning Representations (ICLR)*, 2022.

[25] Yuancheng Xu, Chenghao Deng, Yanchao Sun, Ruijie Zheng, Xiyao Wang, Jieyu Zhao, and Furong Huang. Equal long-term benefit rate: Adapting static fairness notions to sequential decision making. In *AdvML-Frontiers workshop, ICML*, 2023.

[26] Yuancheng Xu, Yanchao Sun, Micah Goldblum, Tom Goldstein, and Furong Huang. Exploring and exploiting decision boundary dynamics for adversarial robustness. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[27] Kaiwen Yang, Yanchao Sun, Jiahao Su, Fengxiang He, Xinmei Tian, Furong Huang, Tianyi Zhou, and Dacheng Tao. Adversarial auto-augment with label preservation: A representation learning principle guided approach. In *Neural Information Processing System (NeurIPS)*, 2022.

[28] Sicheng Zhu, Bang An, and Furong Huang. Understanding the generalization benefit of model invariance from a data perspective. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021.

[29] Sicheng Zhu, Bang An, Furong Huang, and Sanghyun Hong. Learning unforeseen robustness from out-of-distribution data using equivariant domain translator. In *International Conference on Machine Learning (ICML)*. PMLR, 2023.