Statement of Purpose

Pura Peetathawatchai

My academic interests lie in the intersection of **artificial intelligence** and **cybersecurity**. In my research, I treat **AI as a novel attack surface**, the same way the cybersecurity community appoaches hardware, operating systems, and the web from an adversarial lens. My interests encompass both offensive strategies—such as **membership inference**, **training data extraction**, and **data poisoning**—and defensive techniques, including **differential privacy**, **machine unlearning**, and **cryptographic solutions** such as homomorphic encryption. I am also curious about the **future role of AI in the cybersecurity landscape**, particularly in identifying new vulnerabilities and developing new exploits. Pursuing a PhD in Computer Science at Stanford will enable me to delve deeper into these challenges and make meaningful contributions to advancing both the AI and cybersecurity domains.

My pursuits have always been driven by a desire to apply technology for social good, particularly in my home country, Thailand. However, it wasn't until my undergraduate years at Cornell that I began to narrow my focus. Through coursework and internships, I ventured into data science and machine learning, which provided a strong technical foundation. In my final years, after studying and doing research in cryptography, I developed a deep interest in cybersecurity. This curiosity led me to pursue a Master's degree at Stanford, where I sought to further explore the field. My time at Stanford coincided with the surge in generative AI, prompting me to expand my knowledge of machine learning, particularly deep generative models, while simultaneously building a solid foundation in cybersecurity and applied cryptography. As I explored both domains in parallel, I became increasingly fascinated by the intersection of their ideas. These growing interests, combined with a desire to tackle deeper questions in both fields, naturally drew me into academia, where I could pursue these inquiries in a more structured and impactful way.

My first experience with research was at Cornell, where I worked with Professors **Noah Stephens-Davidowitz** and **Huck Bennett** on **post-quantum (lattice-based) cryptography** [1]. We studied the computational hardness of $\mathbb{Z}$SVP (finding the shortest non-zero vector in a rotation of $\mathbb{Z}^n$) and proposed a novel public-key encryption scheme based on this problem. In my first year at Stanford, I was honored to be selected for Professor **Andrew Ng**'s **AI for Climate Change** bootcamp, where I engineered neural networks to attribute methane plumes to landfills and concentrated animal feeding operations (CAFOs) using satellite imagery. I significantly expanded the existing codebase to support multispectral input, semantic segmentation and more granular error analysis. This experience honed my machine learning research skills, including reading research papers, working with key tools such as PyTorch Lightning, Huggingface, and GPU clusters, and adapting to unfamiliar codebases, libraries, and computing environments.

In pursuit of my growing interest in AI safety, I joined Professor **Sanmi Koyejo**'s Stanford Trustworthy AI Research Lab and co-authored a paper [2] accepted to COLM 2024 (28.8% acceptance rate)

investigating the fairness implications of fine-tuning large language models (LLMs) and vision transformer models via **Low-Rank Adaptation** (LoRA) [3]. Our study was the first empirical investigation into the **fairness properties of LoRA**, and we found no consistent evidence that LoRA worsens sub-group fairness compared to full fine-tuning. This experience also gave me hands-on experience in implementing and working with membership inference attacks, particularly the **Likelihood Ratio Attack** (LiRA) [4].

Additionally, under the guidance of Professor **Albert No**, I have been leading research investigating **privacy-preserving style transfer** approaches for **diffusion models**. This opportunity helped me grow as an independent researcher, building confidence in navigating unknowns–knowing when to solve problems on my own and when to seek guidance from mentors. Moreover, I gained the initiative to set my own research directions, identifying key areas for further investigation and designing experiments in light of new data. Initially, we explored using **Universal Guidance** [5] to guide the diffusion process towards an image style without leaking too much information from the target dataset. However, after finding that Universal Guidance did not yield the desired results, I proposed and investigated the potential of using **Textual Inversion** [6], which offers the similar advantage of being lightweight (requiring no model finetuning) while having more established success in the style transfer literature. We are currently drafting a paper, for which I am the first author, for submission to CVPR 2025.

In another line of research, I have been collaborating with Professors **Dan Boneh**, **Daniel E. Ho**, and **Percy Liang** from Stanford to assess the **hacking capabilities and behavior of LLM agents** compared to human hackers, as well as to evaluate the performance enhancements that human hackers can achieve using LLMs. We created *Cybench* [7], a **cybersecurity benchmark** for LLM agents which utilizes professional-level Capture-The-Flag (CTF) tasks, and have submitted it to ICLR 2025. This project gave me extensive hands-on experience in various hacking domains, from cryptography to web security and digital forensics, and a deeper understanding of the tools, techniques, and approaches employed by the hacking community. It also provided valuable insights into how LLMs can enhance, rather than replace, human hackers. To build on this research, we are now working on a follow-up project, *Trinity*, which aims to explore the differences in performance and behavior between human hackers, AI agents, and human hackers using AI assistants.

Having worked with Professors **Dan Boneh**, **Sanmi Koyejo**, and **Percy Liang**, I believe Stanford is the ideal place to continue exploring the multifaceted challenges of AI security and privacy, and ultimately pursuing my goal of contributing to the AI and cybersecurity landscape as a faculty member. I am confident that my unique dual background and deep interest in both domains will bring a cross-disciplinary perspective to my work and enable me to continue making meaningful contributions to Stanford's dynamic community through both teaching and research.

References

[1] Huck Bennett, Atul Ganju, **Pura Peetathawatchai**, and Noah Stephens-Davidowitz. Just how hard are rotations of $\mathbb{Z}^n$? Algorithms and cryptography with the simplest lattice. In *International Conference on the Theory and Applications of Cryptographic Techniques*, 2023.

[2] Zhoujie Ding, Ken Ziyu Liu, **Pura Peetathawatchai**, Berivan Isik, and Sanmi Koyejo. On fairness of low-rank adaptation of large models. In *Conference on Language Modelling*, 2024.

[3] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[4] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914, 2022.

[5] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024.

[6] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022.

[7] Andy K. Zhang, Neil Perry, Riya Dulepet, Eliot Jones, Justin W. Lin, Joey Ji, Celeste Menders, Gashon Hussein, Samantha Liu, Donovan Jasper, **Pura Peetathawatchai**, Ari Glenn, Vikram Sivashankar, Daniel Zamoshchin, Leo Glikbarg, Derek Askaryar, Mike Yang, Teddy Zhang, Rishi Alluri, Nathan Tran, Rinnara Sangpisit, Polycarpos Yiorkadjis, Kenny Osele, Gautham Raghupathi, Dan Boneh, Daniel E. Ho, and Percy Liang. Cybench: A framework for evaluating cybersecurity capabilities and risks of language models, 2024.